AD________________

Award Number: W81XWH-06-1-0291

TITLE: Integration of Anatomic and Pathogenetic Bases for Early Lung Cancer Diagnosis

PRINCIPAL INVESTIGATOR: Wei Qian, Ph.D.

CONTRACTING ORGANIZATION: Moffitt Cancer Center
Tampa FL 33512

REPORT DATE: March 2007

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*
01-03-2007

2. REPORT TYPE
Annual

3. DATES COVERED *(From - To)*
1 Feb 06 – 31 Jan 07

4. TITLE AND SUBTITLE
Integration of Anatomic and Pathogenetic Bases for Early Lung Cancer Diagnosis

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-06-1-0291

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Wei Qian, Ph.D.

E-Mail: qianw@moffitt.usf.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Moffitt Cancer Center
Tampa FL 3351

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES- Original contains colored plates: ALL DTIC reproductions will be in black and white.

14. ABSTRACT
It is widely recognized that the diagnosis of early lung cancer should entail a multimodal approach taking age, clinical history, radiology and cytology/histopathology into consideration. Most radiological assessments should be combined with microscopic analysis of tissue samples to reach a diagnosis. Our proposed new universal computer-aided diagnosis (UCAD) system combining anatomical knowledge based CAD (AK-CAD) and computer-aided pathological diagnosis (CAPD), which provides combined and correlated radiological and cytopathological features in diagnosis of early lung cancers, will give radiologists and pathologists an efficient and automatic tool for their diagnosis. The development of a novel automatic quantitative assessment of radiological-pathological combined and correlated features is proposed here for diagnosis of early lung cancer. The integrated AK-CAD and CAPD system will provide radiologists and pathologists an efficient and automatic tool for their detection and diagnosis of early lung cancers.

15. SUBJECT TERMS
lung cancer, computer-aided radiological diagnosis computer-aided pathological diagnosis.

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	93	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Page

Introduction…………………………………………………………….……..…… 1

Body…………………………………………………………………………….. 2

Key Research Accomplishments………………………………………….……….. 23

Reportable Outcomes…………………………………………………………… 24

Conclusion…………………………………………………………………………. 25

References……………………………………………………………………….. 26

Appendices………………………………………………………………………… 30

1. INTRODUCTION

This annual report is concerning the first year progress for our four-year's Peer Review Grant project. The technical objective of the project is to design a new universal computer-aided diagnosis system for analyzing the anatomic and pathogenetic bases of various radiologic abnormalities seen in diseases of the chest and for radiological-pathologic combined diagnosis of early pulmonary nodules. During this year, our work is divided as two parts: one is on Radiology Features for diagnosis and another one is on Pathology Features for diagnosis, which are described as follows:

PART I: Radiology Features: The radiological features are derived from the analysis of the suspicious nodules and corresponding anatomical structures classified by our universal computer-aided diagnosis (UCAD) system based on helical CT images. The proposed novel UCAD system is named as anatomical knowledge–integrated CAD (AK-CAD) here, which is proposed for obtaining radiological features and for the detection and localization of pulmonary nodules in CT images. It is recognized that, because of the camouflaging structures, the use of gray-level information alone is insufficient to detect lung nodules, especially tiny nodules. In consequence, we will apply the knowledge of anatomical structure to guide feature extraction and feature analysis for nodule detection and diagnosis algorithms. An information sharing and coordinating architecture will be implemented to control the communication between anatomical models and image processing modules for writing, reading, deleting, and updating the information as lung nodules are identified. This AK-CAD system will offer high sensitivity and specificity for the automatic detection of possible lesions, allowing radiologists to focus their time and attention on the most likely regions of cancer cases, with expected decreases in the number of missed diagnoses, so prevalent now, of early stage lung nodules.

PART II: Pathology Features: Pathologic features corresponding to the above detected nodule areas can be obtained for further diagnosis; The pathologic features include cytopathological features and histopathological features, extracted by using our automatic UCAD system respectively based on cytological images from fine-needle aspiration biopsy (FNAB) for cytopathological features and histological images from open biopsies or surgical resections for histopathological features. The proposed novel UCAD system is named computer-aided pathological diagnosis (CAPD) system here, which is proposed to identify different cancer cells in the biopsy images. The proposed method provides quantitative results in diagnosing cancer cells. The new technology, CAPD, is enabling us to identify features of individual cells in ways unimagined by pathologists. For example, all cell types, depending on their functions, have unique, identifiable "signatures" -- special characteristics, such as, which genes are active and what proteins or the cell manufactures other cellular products. Our proposed CAPD enables us to differentiate those signatures. During the transformation of a normal cell to a cancer cell, the signature changes, and the change becomes a signal of the presence of cancer, which can be detected by our proposed CAPD system.

REPORT BODY

PART I

1. Introduction

Lung Cancer is currently the leading cause of cancer deaths in the United States among both men and women, accounting for estimated 3l % of cancer deaths in men and 27% of cancer deaths in women, with a 5-year survival rate of only 15% [1-3]. Meanwhile, the data shows that the 5-year survival rate is much higher if nodules were 20 mm or less in diameter [3]. This finding provides us the benefit and the motivation for early detection of small size lung nodules or lesions that may be cancerous. Hence, it is desirable to find out a method or combination of methods that offers the most accurate and sensitive screening. In terms of screening performance conventional chest radiography has historically shown a low sensitivity for the detection of small lung nodules [4]. Meanwhile, helical computer tomography (CT) system is capable of performing the scan and image reconstruction simultaneously and high-resolution CT has also proved to be effective in characterizing edges of pulmonary nodules [5]. The imaging protocols on single detector scanners typically generate about 40 images in a thoracic CT exam. These large volume data sets are impractical to review in current radiology practice. Even though conventional chest radiograph will remain the norm in clinical practice, its interpretation is an extremely difficult and challenging problem. The variations in interpretations of abnormalities by radiologists still remain unacceptably large. Hence, comprehensive computer-aided diagnosis (CAD) methods are necessary to aid radiologists improve efficiency and accuracy of diagnosis. CAD has shown to be effective in screening mammography and in lung nodule detection [6].

But detecting small nodules non-invasively is not an easy task because of the infinite variety of appearances of nodules on the CT images. Nodules show as relatively low contrast white circular or oval objects with sharp margins within the lung fields, while other structures such as the vertical blood vessel sometimes have similar image characteristics or form superimposed shadows causing high false positive rate. Images that include abnormal anatomical structures also pose a serious problem for lung nodule detection. The traditional image processing techniques typically segment an image into different regions based on intensity characteristics. Typically image segmentation is based on threshold, edge detection and feature-based pixel classification. These systems are effective and useful but do not provide a high-level representation of the image content. Experience has shown that many real world problems are not amenable to simple universally applicable methods unless additional knowledge is included. Few systems have been reported that use anatomical knowledge to perform image interpretation, specifically for the lung nodule segmentation in chest radiography [7-10]. They use constraints on features such as expected size, shape, texture, and relative position to segment regions of interest. Also, anatomical knowledge is embedded within the segmentation algorithms, making it difficult to extend to other problems. In our system, the anatomical knowledge is stated explicitly in a lung anatomical model that is independent of the image processing algorithms. It is used to guide the image processing algorithms to get better result in the advance lung nodule detection system. Additional knowledge, such as anatomical and pathology knowledge that is not derived purely from the image data make the segmentation and classification more adaptive and robust. So, our research objective is to develop such a knowledge-based CAD module for early lung nodule detection on helical CT images

Any individual image processing algorithm has its own strengths and weaknesses when applied to a specific problem. One can develop a mechanism to choose an optimal algorithm or combination of several algorithms according to the nature of the problem. In our system, different algorithms are stored in the algorithm library and they are chosen according to the nature of the specific problem. In a knowledge-based approach there is a need to combine every available knowledge and methods to one highly compacted system. In this preliminary study, the anatomical knowledge and image processing algorithms are combined in a blackboard system. The blackboard model is a conceptual, high-level organization of information and knowledge needed to solve a problem. It is a general architecture for the dynamic control and use of knowledge for incremental problem solving [11].

The organization of the rest of this paper is as following: Section two briefly introduces the anatomy of lung and how the anatomical structures get incorporated into a semantic network, which acts as knowledge base in the nodule detection system. Sections three and four mainly illustrate the knowledge-based lung nodule detection system that is embedded in blackboard architecture and its implementation. Section five presents the preliminary results and analysis, followed by conclusion.

2. Lung Model in Semantic Network

To build up an anatomical knowledge base of lung, we should obtain some knowledge about the anatomy of the lung, its appearance on computer tomography images and the nature of various abnormal structures. The lungs are a pair of organs located within the thorax. The main function of the lungs is gas exchange. There are three lobes in the right lung and two in the left lung. According to the bronchial anatomy, the lung lobes can be sub-segmented to several parts as shown in Figure 1 [12]. The smallest functional unit of the lungs is the bronchopulmonary segment. A single bronchopulmonary segment can be resected when disease is confined to that one segment. The bronchi and blood vessel in each segment are relatively independent to those in adjacent segments. A large structure adjacent to the border of two segments, for example, could indicate the occurrence of a nodule, since normal lung structures generally do not appear near the border [13]. The sub-segments knowledge was used for 3-D localization of lung nodule candidates in the system.

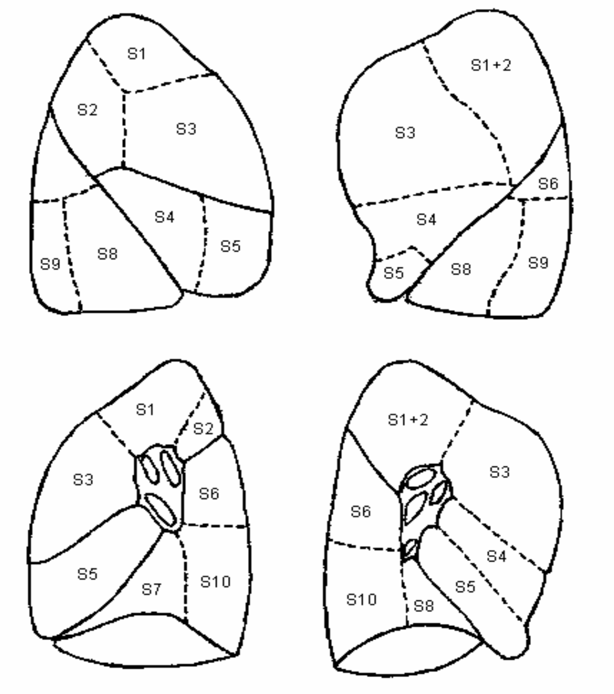


Fig. 1 Bronchopulmonary Segments

In our research, the structure of the lung model is built as a contiguous set of 2-D images of parallel CT slices through the thorax representing the 3-D anatomy. The anatomical knowledge of lungs was used to guide the image and inference processing. We used a semantic network to incorporate the anatomical knowledge into our automatic nodule detection system .The semantic network has the ability to represent the properties of different lung structure and also can represent the interrelationship between different lung parts on each CT slice. The 2-D model describing the structure and features in each slice is organized into a hierarchical semantic network structure where each node represents a lung object and the children nodes are subparts of the object [14]. Fig.2 is a simplified description of the 2-D lung module for each slice. Both terminal and intermediate nodes which represent lung structures are defined by features which describe the object. The nodes are used later during the recognition process to match lung objects against image objects. The spatial relationship constraints are defined only among the nodes stemming from the same parent in the hierarchy (see dashed lines in Fig.2).

At each level of the hierarchy, a 2-D object instantiation is described in several terms, generally, at higher level, which mostly includes large areas, such as background, lung wall, thorax, fewer terms are used because these structures are relatively easy to be segmented, as the level moves down, the objects are confined in a smaller area and they are more likely have similar image property, more and more terms are included into the nodes to deal with the similarity. For a specific candidate that we want to classify as a final result, we need more details about the object to make a successful classification.

The features we used in each level are as follows: (1) Gray value domain, area, and the mean value of gray level at the first level, which includes chest wall, background, and thorax, (2) Gray value domain, area, mean value of gray level, variance of gray-level, cavitation, etc at the second level, which includes hilum, central bronchia tree, and lung field, and (3) Gray value domain, area, mean value of gray level, variance of gray-level, circularity, sharpness of margins, variance of gradient, distance from lung wall, angularity of margins, and orientation at the third level, which includes blood vessels and abnormal structures (nodules).

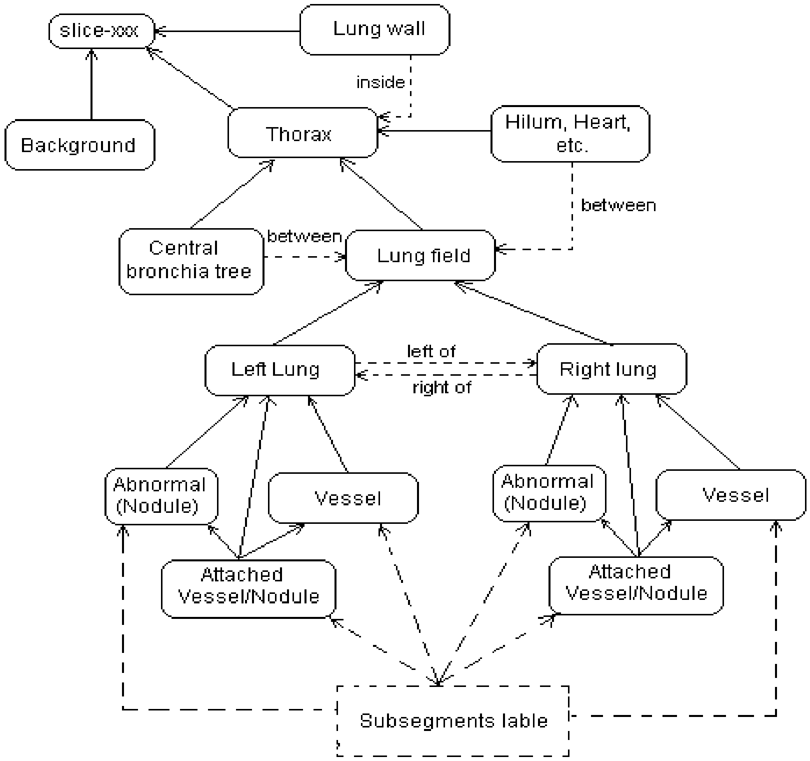


Fig. 2 Semantic network

The dashed lines and "sub-segments label" module are used for 3-D object interpretation. The links to the 2-D instantiations by the "sub-segments label" indicate in which slices 2-D cuts of a particular 3-D structure are typically perceived. They also indicate which context information can be shared among different slices. The sub-segments labeling only applies to the terminal level of the model. In this level, all the candidates in the lung field either are nodule or blood vessel and each candidate will be associated with a specific sub-segment. In our automatic detection system, if a candidate cannot be classified by its 2-D features, a further 3D validation will be applied. The sub-segments labeling helps the system to search and locate the relevant 3-D structures on different slices.

For high-level image analysis, a high-level representation must be derived. In our approach, the interpretation involves matching low-level representation described by numerical parameter to high-level representation using fuzzy sets. Fuzzy set theory allows representing uncertainty and vagueness mathematically using formalized tools that can deal with imprecision [15]. Since knowledge can be expressed in a more natural way by using fuzzy sets, the lung anatomical knowledge representation can be greatly simplified. Anatomical feature descriptions are imprecise and so linguistic variables such as large, medium, and small easily lend themselves in capturing the range of numeric value associated with anatomical features. Fuzzy sets are used to accomplish the transformation of numerical feature values to symbolic descriptions. With the following example, we try to illustrate the use of fuzzy confidence function for the relationship between a suspicious candidate's roundness and the possibility to be a nodule or blood vessel on one slice. The object is assigned a grade of confidence according to its roundness, a grade of 1.0 indicates that it have very high probability to be certain object, 0.0 indicates very low probability, values in between indicate different probability or degree of roundness.

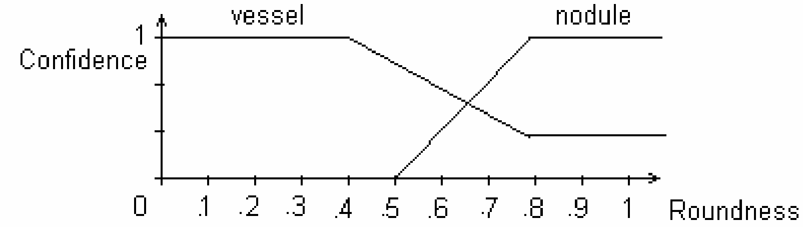


Fig.3. Fuzzy confidence function of roundness

From Figure 3, we can see that if an object has a high degree of roundness, it is more likely to be a nodule and as the roundness reduces, the probability of being a nodule also reduces. Because a nodule generally has a round appearance, we can say with high degree of confidence that it is not a nodule but a blood vessel when the roundness falls below 0.5. But, because of the existence of vertical blood vessels, we cannot clearly decide whether an object with a high degree of roundness is a vessel or not and hence additional knowledge is needed for accurate decision. Other medical knowledge about the features of both the lung nodules and blood vessels are embedded into the lung model by fuzzy set [16]. For example, the shape of the lung nodule is generally circular in the cross-sectional images; on the contrary, the shape of a blood vessel running parallel to the sliced image is generally oblong. The thickness of the blood vessels becomes smaller near the lung wall; however, the lung nodules are generally larger in comparison. Since the peripheral blood vessels are too small to be seen in the helical CT images, shadows along the lung wall are generally nodules or partial volume artifacts. The pixel values (intensity) of the cancer region are uniform. The characteristics of normal structures such as blood vessels depend on their location in the lungs. The vessels in the middle lung region tend to be large and intersect slices at oblique angles. The vessels in the upper lung regions are usually smaller and tend to intersect the slices more vertically. In general, the vessels are densely distributed near the hilum and spread out towards the periphery.

Based on this anatomical information, we define different fuzzy confidence function for different features for each nodule candidate in each slice. It provides an intuitive means of modeling anatomical variability and allows the model to impose soft constraints on the segmentation and matching. An overall confidence score is derived by a composite constraint function, defined as a normalized weighted average of the individual confidence score associated to the object's feature. It is given by the following equation:

$$S_{overall} = \sum_{i=1}^{M} w_i F_i$$

(1) where, wi is the weight of a specific feature, Fi is the confidence score of the feature, and M is the number of features.

3. System Architecture

Some computer vision systems using blackboard approach for communication and control between the different system components have been developed [10, 17]. In this study, the anatomical and image processing knowledge are embedded in a blackboard system [18]. The blackboard model based system architecture (see Figure 4) includes several modules: (1) Blackboard (BB): the central workspace, a temporary database that is continually constructed and updated along the inference process. The dashed line represents the abstraction hierarchy of inference process, with the original image input at the lowest lever and a mapping report at the highest. (2) The lung model knowledge base: it stores the semantic networks described in section two, which contains the descriptions of lung objects, such as their image characters and relationships. (3) The operator scheduler (OS), which contains lists of different image processing routines for each specific problem. (4) The information extractor (IE), which acts on BB for feature extraction and feature conversion. It extracts information from image board and sends it to feature board for further processing. (5) The image processing library, which contains different image processing routines that can be invoked by OS and IE. (6) The inference engine, which collects information from the feature board and lung knowledge source, matches the candidates to the best fit model, that is, matching image objects to lung structures stored in the lung knowledge source. (7) The temporary storage which stores description for each nodule candidate used for 3D validation for more accurate classification.

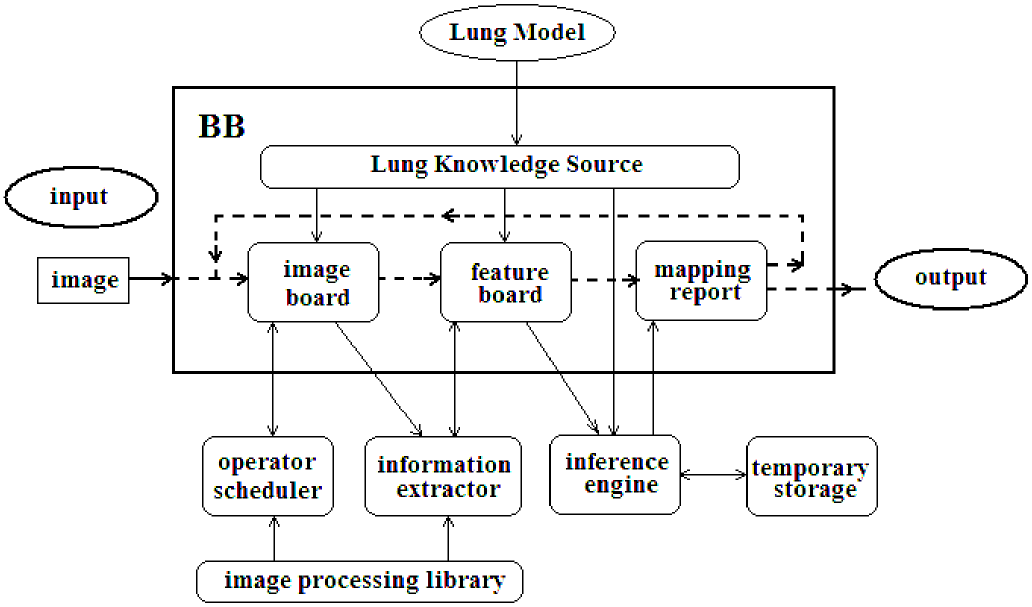


Fig.4. System architecture

In our system, the lung model knowledge base represents a radiologist, OS, IE and image processing library together represent an image processing expert. In each step in the inference processing on blackboard, the radiologist and image processing experts contribute their expertise to guide the processing. The function of inference engine is to map a candidate extracted from the image to a real anatomical structure stored in lung knowledge base.

The blackboard workspace is arranged in a hierarchy of three levels. They are, in order of increasing abstraction. (1) Image board: all images that include original image, processed image, temporary image, etc., are stored in the image board. Lung knowledge source and OS work together on the image board. From the original input image, they produce necessary subsequent images for further processing. (2) Feature board: It is the place where feature description of each image region or candidate is stored. The descriptions derived from the image board by IE are embedded in a frame based structure. (3) Mapping report: This module is a report space that contains partial and final mapping result for each candidate. (4) Lung knowledge source (KS): the lung description information that is input from the lung model is stored in this module, which is involved in every inference step. Lung KS provides information about what kind of features should be extracted for a specific case in each step.

4. Implementation

The system operates in “step”, on each step, the radiologist and image processing expert work together to guide the inference, that means the lung model and image processing modules exchange necessary information through the blackboard. Lung model provides information about what kind of image candidates should be segmented, what kind of features should be extracted, etc. According to the information that is post on blackboard, image process modules operate on images. According to the extracted features, each image candidate is matched to a specific lung anatomical structure that is modeled in lung KS by inference engine. Because the lung model is organized in hierarchical structure, for each level, the system runs through the whole processing, until it reach the terminal level, which includes nodule candidates we are interested in. These classified candidates are the final system output. The nodule candidates are stored in temporary storage. If a nodule candidate cannot be classified by 2-D features on single slice, the inference engine accesses its corresponding candidates on adjacent slices to make a 3-D validation. The sub-segments labeling helps the inference engine search the corresponding candidates on different slices.

A. Work on Image Board

The first step is to input the original CT image to the image board. In our preliminary study, for each input image, we manually set a slice number that associates with the position in lung model. According to the slice number, blackboard accesses the lung model and save the corresponding lung semantic network in lung knowledge source. By collecting information from semantic networks, the operator scheduler chooses different algorithms and parameters, similar to the automatic selection of image processing algorithms that has been described by a system called Borg [19]. For each level of the hierarchy in the semantic networks, we set a sequence of image processing according to our stand-alone trial and error experiments. The algorithm lists are stored in operator scheduler and when a certain level is scheduled for segmentation, image board invokes the sequence of algorithm from the list.

The current system with the hierarchical strategy starts with a segmentation step to separate image objects: background and thorax (lung field). Because the gray value on

CT scans have a well-defined physical meaning in terms attenuation coefficient of biological tissues for X-rays and the gray value of the hilum and lung walls are much higher than that of the air (background) surrounding the patient, the segmentation is based on a simple global gray value threshold. This method successfully separates the surrounding air region and thorax for all cases in our data. After this, IE and inference engine classify each region according to the knowledge stored in lung KS, then the system moves to the next level. At this level, the processing will focus on a finer subdivision of the lung. However, the finer region often corresponds to more than one object and usually these objects have similar image properties, so more complex segmentation algorithms have to be used to split it into smaller segments. A k-means clustering technique [20] was then used to automatically segment a CT slice into lung field and regions containing the lung wall and mediastinum. Some of the nodules adjacent to lung wall may be excluded from the extracted lung field. To make up these impairments, we used a method described in [21], where the curvatures of the lung border were calculated and the border was corrected at locations of sharp curvature change by straight lines. At the final level, nodule candidate segmentation in the lung field, a fuzzy C-means algorithm [22-23] was applied. With a finite set of features X = {x1, x2,.., xm}and the number of cluster centers {vk, k=1,2,..,C}to be calculated, the assignment of the m features to the C clusters is represented by the proximity matrix which expresses the fuzzy affiliation of feature xi to the cluster center vk.

$$U = [u_{ik}], u_{ik} \in [0,1], i = 1,2,...m; k = 1,2,..,C \qquad (2)$$

One problem after the lung nodule candidate segmentation is that different structures within lung can merge into a single connected region. The result is multiple objects merged into a single larger object. In order to reduce the distortion due to merging, a binary splitting is performed on the detected objects as described in [24].

B. Feature Extraction

The next step on the blackboard is feature extraction. Lung KS posts what kind of features should be extracted. IE access the feature board and extract different features from the image board for each candidate and post these features back to feature board for further inference. In the current system, a candidate's features are stored on feature board in a frame-like manner, one frame for one candidate. The frame encapsulates all the information related to a candidate into one package, thereby significantly simplifying the data flow and control on blackboard. The features obtained by IE include: 1) Shape and size features, such as area, perimeter, width, height, roundness, smoothness of margin, orientation. 2) Statistic features, such as mean gray value, standard deviation of gray value, minimum and maximum gray values. 3) Location features: distance from border/structures, centroid, locations in sub-segments, spatial relationship with other regions, etc.

C. Inference Engine

The main function of the inference engine is to classify image candidates to lung structure. It also performs a number of tasks related to decision-making within the system. After feature extraction, inference engine gathers information from lung KS and feature board for matching. In order to allow high-level interpretation process, some of features in lung model, such as size, position, roundness, are defined by fuzzy set. IE converts some features to symbolic term according to the fuzzy set stored in lung KS. The interpretation of lung CT implies finding an acceptable mapping between image and objects in each level.

D. 3D Validation

If a candidate cannot be classified within a 2D slice, then 3D relationship between objects on adjacent slices has to be used for classification of lung nodule candidates. In the analysis of 3D objects on CT images it is desirable to detect the relative position of interested anatomical structures on different slices. One way to achieve this is to register the image to a model of the depicted structure. Based on the position of the nodule candidate, we associate the nodule candidates to specific sub-segments on each slice and assign coordinates to each candidate within the sub-segments. Figure 5 shows different segments on different slice.

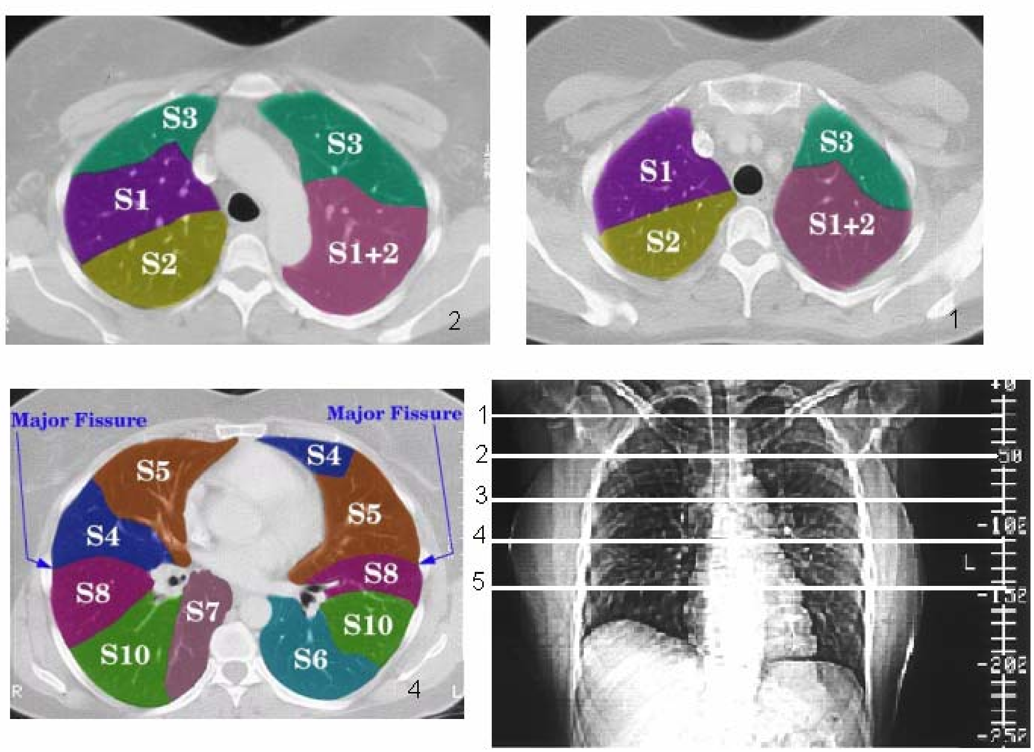


Fig.5 Segments on different slice

The sub-segment label and coordinates of each nodule candidate are stored in the candidate's frame in temporary storage. When the system wants to access 3D related candidates for a specific candidate, first the system uses sub-segment label to find all the candidates in the same segment on adjacent slice, then use coordinates to calculate the distance between two candidates. The distance is used as criteria to choose candidates that have high possibility of 3D relationship. Here, the segmentation was carried out in 2D while 3D information was included after the segmentation for classification. True 3D segmentation may provide more accurate shapes of 3D structures. In this system, blood vessel reduction in the first inference step partially eliminates a large number of vessel structures during the 2D analysis before forming a 3D object. Once 3D objects were formed, the following features were extracted for each 3D object: volume, surface area, average gray value, standard deviation, sphericity. The volume was calculated by counting the number of voxels within the object and multiplying this by the unit volume of a voxel. For surface area calculation, the voxels along the surface of the object were considered. Future extension of the 3D features will include 3D shape description such as, smoothness of the surface and skewing. These features may provide a better 3D perspective of the objects and thus improve differentiation of nodules and normal structures.

5. Preliminary Results and Analysis

Anatomical structures were segmented and nodule candidates were classified from CT images automatically by the proposed system. The CT images were acquired using LightSpeed (GE Medical Systems) scanners with a slice thickness of 8 mm. Each image is 512x512 pixels and each pixel is about 0.4x0.4 mm2 represented by 16-bit gray level. To evaluate the performance of the automatic lung nodule detection system, the system was applied to 28 helical CT images of 5 patients. The test images were not chosen randomly, but were selected with the aim that most of the detectable abnormalities were covered.

Figure 6 shows an example of step by step processing of CT images in our system. Figure 6(a) is the original CT image while Figure 6(b) shows the binary image after thresholding. The threshold value in different slices ranges from -300HU to -550HU for segmenting the image into different areas: thorax, lung field and the background. According to the size and mean gray value of each segmented region, the inference engine uses predefined rules stored in the lung knowledge model to classify and remove the background as shown in Figure 6(c). In the hierarchical process, the steps from the second level on will not be applied on the background. A k-means clustering was then applied to the segmented image. K=2 with one cluster for the lung field and the other for the soft tissues and bones. Only pixel gray values were used in. clustering. In the detected left and right lungs, we show only the right lung for further illustration of our results for brevity. Figure 6(d) is the segmented right lung field classified by the system which contains some impairment including interior “cavities” due to the existence of large transitions in gray values of pixels around blood vessels and concave boundaries. The system computes the cavity size as the number of pixels in the area and those cavities that are smaller than a predefined value are filled up. Lung nodules adjacent to the lung wall may be excluded from the extracted lung field. To alleviate this problem, the curvature of the lung boundary is corrected at locations of sharp curvature changes by straight lines. The repaired lung field and then superimposed on the original image are shown in Figures 6(e) and (f). For the lung nodule candidate segmentation, a histogram based fuzzy C-means clustering algorithm with Euclidean distance measure was applied with the number of clusters C=2 and fuzziness hedge m=2. Figure 6(g) shows the result of clustering based only on a single feature of pixel gray values. Since multiple objects tend to merge into a single large object as shown in the figure, we use the splitting technique to separate the connected nodule candidates as shown in Figure 6(h). Features were extracted for each nodule candidate in the segmented regions and then the inference engine by applying the classification rules finds the best match according to the overall confidence scores based on the composite confidence function given by

$$S_{overall} = \sum_{i=1}^{6} w_i F_i = 0.4F_{round} + 0.1F_{size} + 0.1F_{mean} + 0.1F_{\mathrm{var}} + 0.15F_{m\arg in} + 0.15F_{dist} \quad (3)$$

Where, Fround is the roundness, Fsize is the size, Fmean is the mean of gray values, Fvar is the variance of gray values, Fmargin is the smoothness of margin of nodule candidate, and Fdist is the distance of lung wall. Each nodule candidate has two overall scores; one for nodule and the other for the blood vessel. In this research, the weights (wi) of the composite constraint function were determined empirically with guidance from our trial-and-error experiments and validated by domain experts. The fuzzy confidence score for individual features (Fi) is decided by the fuzzy set stored in the lung knowledge base. The fuzzy membership functions for each feature in the model are trapezoidal in nature as shown in Figure 3. Figures 6 (i) and (j) show two different results by selecting different thresholds for overall score to classify as blood vessels, lung nodule candidates, and unclassified candidates. By changing the weights and features combined in composite confidence function, this system imposes soft constraints on nodule detection. The classification result is marked out on the original image in Figure 6(k).

Anatomical structures were segmented automatically from CT image using the system described. These computer-marked abnormalities were compared against the truth files in which the nodules were identified by a radiologist. The system performance was summarized in Table1, where a nodule is considered as a positive result. The nodule detection system successfully detected most suspicious regions from the chest CT images. There were a total of 1069 nodule candidates in the test set with an average of 38.2 (1069/28) 2D objects per image. In the 2D domain classification, the detection sensitivity was 94% (46/49). We used relatively low threshold to classify the nodule candidates to achieve high accuracy, the tradeoff was a high false positive (FP) rate of 3.6 per image (96/28).

Table 1 Classification results of nodule candidates

	2D domain Classification		2D+3D domain classification	
	TRUE	FALSE	TRUE	FALSE
NEGATIVE	924	3	998	6
POSITIVE	46	96	43	22
TOTAL	970	99	1041	28

Features based on small 2D structures may not represent the general characteristics of a nodule. For example, the top and bottom slices of a 3D object might be very small. So in the current system, the candidates classified by 2D features would not be further considered after the classification and hence the false negative (misclassified nodules) in 2D+3D domain is higher than that in 2D domain. A better strategy would be to change the classification previously decided by 2D features according to the 3D object after forming a 3D object. In the 2D+3D domain, the detection sensitivity is reduced to 88% (43/49) but the FP rate improved significantly to 0.79 per image (22/28).

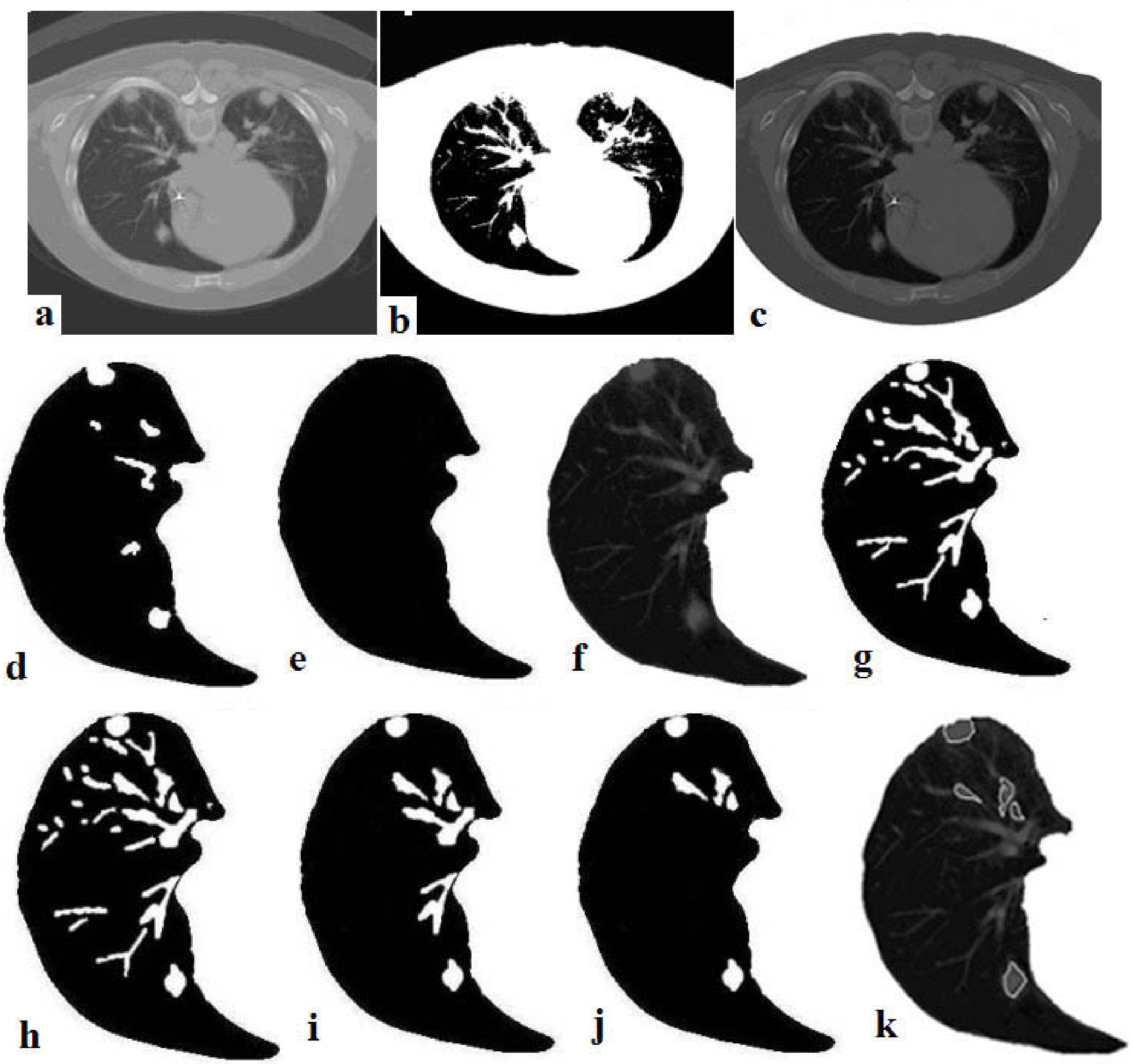


Fig. 6 Experimental results of our nodule detection system (shown right lung only)

The preliminary experimental results obtained were encouraging demonstrating the proof of concept and illustrating the potential of the automated knowledge-based lung nodule detection system. Given the simplicity of the current model, the number of abnormalities that could be identified was encouraging. The design of the system was motivated by a desire to develop a general-purpose framework for knowledge-based medical image interpretation. By adapting to different knowledge source, the system is potentially useful in other applications. However, to be clinically useful, the system would need to be expanded to include more subtle structures. The introduction of anatomical knowledge increased the ability of an automated system to discriminate between different structures, particularly those similar to X-ray attenuation. At current stage, the features and weights in confidence functions were selected by experienced image processing experts and radiologists. More specific features and classification criteria tailored to the different regions will also need to be designed to better exploit the different characteristics of nodules and normal structures. Fuzzy classification rules in the current knowledge-based detection system may better accommodate the large variations in the features of the nodules and normal structures and may lead to further improvement in the recognition. However, the design of a more complex classification rules and systems has to be based on a larger data set. Further work with a larger data set will be required to design the individual and composite confidence functions.

PART II

1. Introduction: Clinical lung cancer diagnosis depends upon the pathologist's interpretation of morphological features of histological and cytological specimens. These interpretations may be assisted by detection of cancer cell related moleculars to improve sensitivity and specificity of the diagnostic techniques. Imaging technology provides an objective way to quantitative analysis of tumor cells morphology. Current imaging practices are mostly manual, time-consuming, and tedious, yielding subjective and imprecise results. In order to improve the situation, many methods for computer-aided diagnosis of cell images have been designed [25 – 34]. The methods including common commercial tools for computer aided morphologic image analysis use region-based methods and threshold based methods. Region-based methods separate object from background by region growing, region splitting, and merging of regions to segment the image. Threshold-based segmentation is a simpler method based on single pixel classification; a feature value such as gray level is associated with each pixel; this value is compared to the threshold to classify a pixel as an object or background. Determination of the threshold is critical, a simple method is to select the threshold as determined by histogram. More sophisticated versions of this are given in Refs. [35, 36]. The problem with these approaches is that they employ only local (single pixel) information. Cellular images share the following characteristics:

- Poor contrast: Object (cell) gray levels may be close to that of background.
- Many cluttered objects in a single scene: A high number of occluding objects make image segmentation difficult.
- Low quality: Traditional staining techniques introduce a lot of inhomogeneity into the images, where not all of the parts of the same tissue are equally stained.

Our CACD system accurately identifies cell features by adjusting the parameters of CACD modules based on an advanced adaptive strategy and muti-resolution/multi-orientation techniques.

Lung cancer is currently the leading cause of cancer death in the US among both men and women, accounting for 3l% of cancer deaths in men and 25% of cancer deaths in women. The 5-year survival rate for lung cancer is only 13%. An important reason for the high death rate and low survival rate is that lung cancer often is detected only after it has already metastasized [33, 39, 40] and almost all lung cancers have spread beyond the primary site at the time of diagnosis. Studies have shown, however, that the 5-year survival rate for patients with a peripheral squamous nodule 20 mm or less in diameter is as high as 88%, which indicated the benefit of detecting lesions early when they are still small. It is desirable, therefore, to identify a method or combination of methods that offers the most accurate and sensitive screening for small nodules. This is in the selected topic area of lung cancer screening and has special importance for military and public health outcomes, such as smokers and former smokers, who will always have an elevated risk of developing lung cancer, and stand to benefit from the development of any kind of screening method.

Imaging protocols on single detector scanners based on helical CT technique typically generate about 40 images in a thoracic CT exam, multislice protocols may generate 300-600 high-resolution axial images. These large data sets are impractical to review in current radiology practice, and thus efficient methods of image interpretation are required, such as computer-assisted detection (CAD) [31, 32, 33, 34]. Clinical image analysis has also revealed large inter- and intra-patient variations and considerable overlap in attenuation value distribution both among different nodules and between nodules and normal tissues, indicating that CAD methods based on image contrast information alone may be insufficient for successful differentiation between normal and tumor tissues [35, 36, 37, 38, 39, 40]. The anatomical knowledge structure has advantages in tissue differentiation through robust feature description [41]. Proposed here is development of an anatomical knowledge (AK) integrated CAD method that would offer a high sensitivity of automatic detection of possible lesions and allow radiologists to focus their time and attention on the most likely cancer cases.

In general, measures for early stage lung cancer diagnosis mainly include those utilizing X-ray chest films, CT, MRI, isotope, bronchoscopy, etc, among which the most definitive measure is the pathological tissue diagnosis on the specimens of needle biopsies from the detected nodule areas. At present, experienced pathologists are usually required to analyze the specimens of needle biopsies for diagnosis. But, some cancer cells are a diagnostic challenge and may be misinterpreted during screening. Our proposed CAPD system will accurately identify the cell features by adjusting the parameters of CAPD modules based on our advanced adaptive strategy and muti-resolution/multi-orientation techniques, which can track the abnormal changes of even one cell and can distinguish the cancer cells from those normal ones when they still look "normal".

Finally, Expert Radiology-Pathology Panel will review each detected suspicious cancer case, from CT images to histological/cytological images, including the combined diagnostic results produced by UCAD system.

2. MATERIAL AND METHODS

2.1. Cell Lines

To train the CACD system to assess features of lung cancer cells, we evaluated the morphologic features of culture lung cancer cells. The lung adenocarcinoma cell line A549 was obtained from the American Type Culture Collection, Rockville, MD. We established the other cell lines from primary culture of resected lung cancer. Cells were grown in RPMI media supplemented with 10% heat inactivated fetal bovine serum (R10), 1 mM glutamine, and antibiotics, and passaged weekly at subconfluence after trypsinisation. RPMI 1640, trypsin and sera were obtained from Mediatech, Inc., Herndon, VA. Cultures were maintained in humidified incubators at 37ºC in an atmosphere of 5% CO2 in air. For morphology and immunohistochemistry studies, adherent cells were trypsinized and then preserved in PreservCyt (CYTYC Corp., Boxborough, MA), and fixed cytospin preparations were stained with hematoxylin-eosin and for γ-H2AX staining with PX technique (as described below).

2.2. Immunohistochemistry

Formalin fixed, paraffin embedded tissue from adenocarcinoma and squamous cell carcinoma of the lung were obtained from the Moffitt Tissue Procurement Core and cut into at 3 µm sections. Standard IHC with antigen retrieval was performed using a primary antibody to γ-H2AX (Rabbit, polyclonal IgG from Upstate Biotechnology, Lake Placid, NY) applied at 1:700 dilution, overnight at 4ºC in a humid chamber. IHC was completed on the DAKO autostainer using VectorElite – PX Rabbit detection and DAB chromogen.

2.3. Feature Description of Cancer/Normal cells

If we define the differences between normal cells and cancer cells as three levels—histology, individual cell, and nuclear of cells as described in [37-44], the computational features and corresponding
pathological description can be summarized in Table I.

Table I

Difference between Normal Cells and Cancer Cells

Features	Pathological Description
Shape: measured by area, curvature, boundary, first and second deviations.	Cancer cells usually have rounded up shape and can maintain their rough spherical surface.
Confluence: measured by contrast, affined invariant moments and compactness.	Normal cells usually aligned in the same direction, and highly confluent with only a few gaps between the cells. As far as cancer cells concerned, they always keep division, so when there is no room the cancer cells start to grow on top on one another creating an amorphous cell mass.
Size and Amount: measured by area, curvature, boundary, energy and perimeter length over cell.	Normal cells varied in size, from 80 to 350gm. Cancer cells are predominantly small, normally varied from 45 to 85gm. Cancer cells were often closely related in clusters, suggesting that cell division is occurring.
Ratio of N/C: measured by solidity, energy, intensity variation and areas.	Cancer cells have relative bigger ratio of nuclear to cytoplasm than normal cells
Inhibition: measured by co-	Very important difference of normal cell and cancer cell. Normal cell won't go on their divisions when they contact

occurrence matrix, similarity, circumference and the sum and difference of histogram	each other, but cancer cells keep dividing in all the space they could occupy, then produce a great amount of small, high density, useless cells

3. ALGORITHM DESIGN FOR CACD SYSTEM

The CACD system includes the following basic stages: (1) preprocessing algorithms using adaptive strategy for AFWF, TSF, DWT and TSWT modules for enhancement of cellular images to get the histological characteristic features, (2) declustering for isolation of cell and nuclei using fast distance-transform to get individual cell characteristic features and the nucleolus features, (3) the image segmentation using an unsupervised Hopfield artificial neural network classifier and the labeling of the segmented image based on chromaticity features and histogram analysis of the RGB color space components of the raw image, (4) feature description, extraction and classification to finally identify whether a cell is normal with high confidence, and then to deal with the cell that are judged as cancer cells with cancer subtype classifications.

3.1. Modification of preprocessing CACD modules using the adaptive strategy

Adaptive fragmentary window filter (AFWF): The AFWF, originally designed to detect circular patterns in a digitized radiographic image [45], was modified to detect digital cellular images in a histological section by analyzing the edge gradient of all pixels within a small window. This window is chosen to be at least as large as the smallest cells to be detected. Since the majority of cells display circular geometry, analysis of the edge gradients allows identification of the cells center as the locus of a sufficient (chosen) number of edge gradient vectors. The nodule boundaries are found by employing a circularity template, which is a cross-correlation of the individual normalized vector components. The measuring of locality is used to isolate the nodule by eliminating detected gradient vectors with components that are unequal beyond a chosen threshold. This threshold allows the detection of nodules that depart from strict circularity. Finally, the resulting image is then subjected to a spatial and multi-scale analysis to isolate suspicious areas as local intensity maxims. The circularity is also an important feature to identify normal and cancer cells. The AFWF was modified to aid the cytological diagnosis.

Adaptive TSF module: The advantage of the current tree-structured nonlinear filtering (TSF) for image noise suppression is that its application does not require a priori knowledge of the local statistics within the filter window; i.e., it is nonadaptive and therefore computationally efficient [46]. Although the TSF has already demonstrated good performance, adaptive methods will be explored as an optimization strategy. **Adaptive criteria:** These methods will be included: (1) develop an adaptive technique for automatic parameter selection for the TSF, i.e., parameters K_1, K_2 and K_3 as defined in Equations 7-10 of [46] and (2) develop an adaptive method for selecting the filter window sizes (i.e., from 3 x 3 to 7 x 7) depending on requirements for image detail preservation. **Evaluation criteria:** The initial physical performance of the adaptive filter will initially be evaluated by standard signal processing criteria: (1) localized metrics for noise evaluation such as the normalized mean square error (NMSE) and difference images to show structured noise; and (2) inclusion of the effect of application of the directional wavelet transform (DWT) module to the same simulated images to evaluate possible artifact generation.

Adaptive DWT module: This module was designed as a bank of wavelet filters implemented by using adaptive combiners with different weight factors [47]. It can, therefore, be uniquely modified for higher order N directional filters. For example, a higher order wavelet orientation (N=16) was recently implemented, affecting the direction angle Q_i, i.e., the directional bandwidth of the wavelet functions to allow more selective extraction of directional features. Further improvements in robustness may be achieved through adaptivity of the weights W_i applied to the directional features [47,48]. We have performed an initial evaluation of N=16 and found that it improves preservation of the shape of segmented cell areas. **Adaptive criteria:** We propose adaptive selection of N for each pixel point to match these changes. By adaptively selecting N, the sampling problem is potentially improved for detection of morphology of cell shapes. Alternatively, for the higher order N, an improvement in the signal/noise ratio for detection of morphology of cell shapes. The range of N, which influences the angular bandwidth frequency and directional sensitivity, will be adaptively selected within 4-32, which corresponds to a 45° - 5.63° arc width, as described by [48, 31] and provides higher sampling and more mathematically rigorous method. **Evaluation criteria:** The physical performance of the adaptive method will initially be evaluated by using the different cell examples, which contains linear and other structures, to determine if appropriate structures are identified or any artifacts generated.

3.2. Declustering for isolation of touching cells and nuclei

A newly developed distance-transform is proposed for separating a cluster of cytological cells into individual cells with clear contours of cells' boundaries, and is also designed to locate and describe the nuclei inside the cells. The distance-transform yields the minimal distance to the boundary of the object for every pixel in the transform's input image **I**(x; y) as represented in Eq. 1 and shown in Figure 1.

$$\mathbf{Y}(x,y) = \{\min_{\forall k} \| (x,y) - \eta(x_k,y_k) \| \} \qquad (1)$$

algorithm (see Figure 1). The area A (e.g. a cell cluster) which is obtained as a two-level grey-image is initialized such that pixels x $\in$ A are assigned the value 1, background pixels are set to zero; then the structuring element given in Figure 1 is moved over the whole image **I**(x; y). If all elements in the image below the structuring element take
With the result of the distance-transform **Y**(x; y), the coordinate of every pixel x = (x; y) in a uniform area (x; y) $\in$A. $\boldsymbol{\eta}(x_k; y_k)$ is the surrounding curve of the area. The distance transform can be approximated using a 5 x 5 structuring element in an iterative values **i**(x; y) ≥ *a* (with *a* = 1 in the first step) then the point below the center of the structuring element is incremented (on the same image). Finally, we first increment *a* and repeat the process.

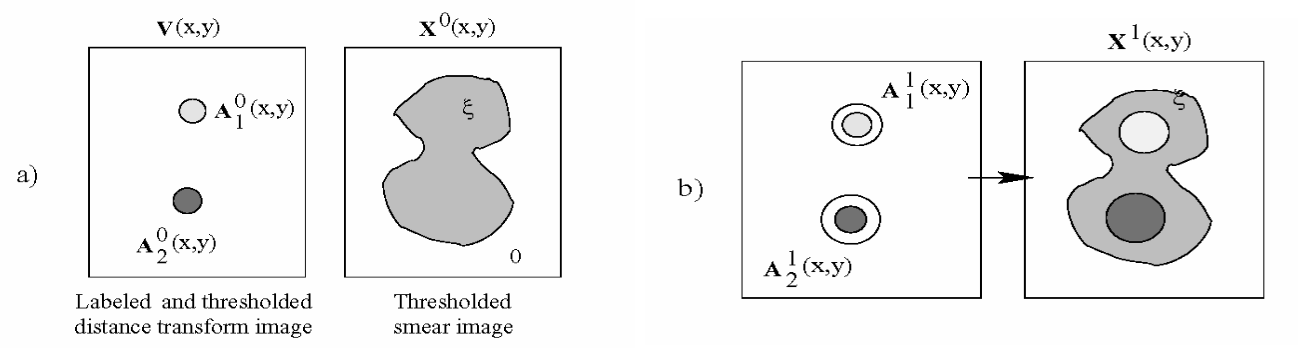


Figure 3: a) Input preparation for the declustering algorithm, b)The first iteration of the declustering algorithm

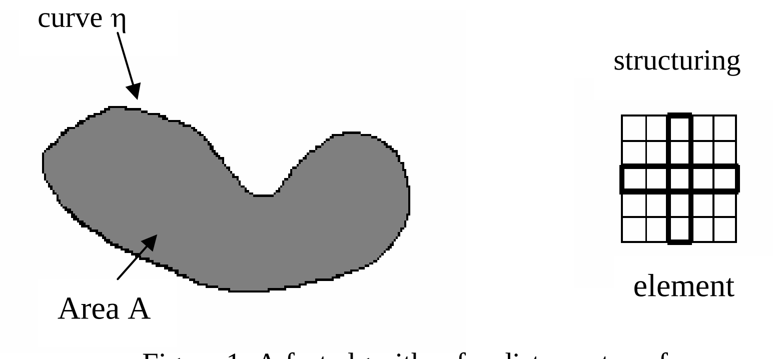


Figure 1: A fast algorithm for distance transform

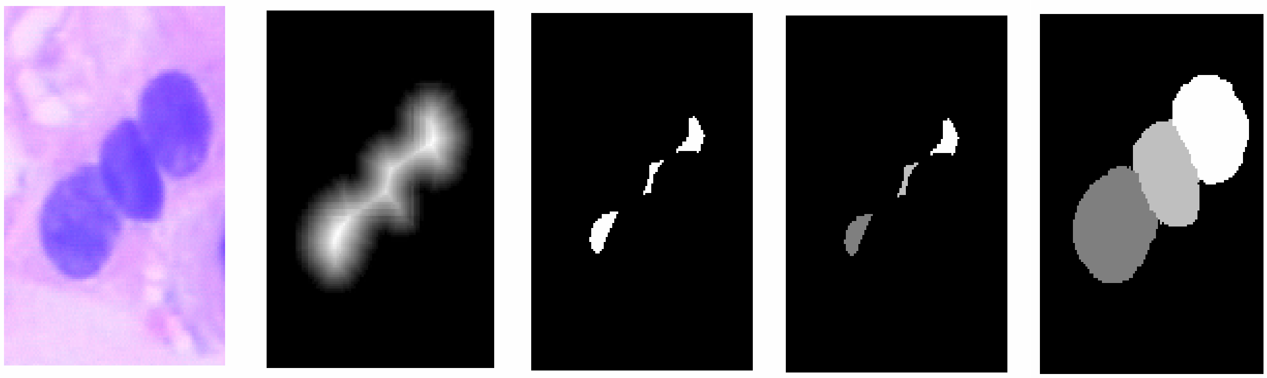


Figure 2: Three clustered cells (from left to right): original image, results from distance transform, threshold over distance transform, labeled threshold, and labeled masks of cells

The distance-transformed image is then thresholded (see Figure 2).

The region growing algorithm is used, which must be controlled by a condition preventing regions from growing into each other (e.g. the regions given in Figure 2 would grow together without such a condition). To ensure that the region growing stops at the boundary of the cell, we first label the thresholded distance-transform image and label the cluster and background: the area of the clustered object is labeled with a value ξ, the background is labeled 0. For the initial areas A_i^0 we obtain (the subscript denotes the label-index of the area, the superscript gives the iteration step), **V**(x; y) is the labeled thresholded distance-transformed image (see Fig. 2 and Figure 3) which consists of regions denoted the labels i. X(x; y) is the pre-segmented image:

$$\begin{aligned}\mathbf{A}_i^0(x,y) &= (\mathbf{V}(x,y)|\mathbf{V}(x,y)=i)\\ \mathbf{A}_i^{k+1}(x,y) &= (\mathbf{A}_i^k \oplus \mathbf{S}) \wedge [(\mathbf{X}^k(x,y)=\xi) \vee (\mathbf{X}^k(x,y)=i)]\end{aligned} \qquad (2)$$

The dilation ($A_i^k \oplus S$) is performed w.r.t the areas labeled i. The iteration is stopped if the

areas A_i^k comprise the whole cluster. The overlapping viz. ($A_i^k \cap A_j^k$) ≠0, *j ≠ i* of the areas is avoided using an extended condition for growing (see Eq. (2)) and by introducing results which were already obtained into the source-image X(x; y) as follows:

$$\mathbf{X}^{k+1}(x,y) = \bigvee_{i=1}^{N} [(\mathbf{X}^k(x,y)|\mathbf{A}_i^{k+1}(x,y)=0) \vee (\mathbf{A}_i^{k+1}(x,y)|\mathbf{A}_i^{k+1}(x,y)>0)] \qquad (3)$$

The I-th region comprises the whole cell, if $\sum_{i=0}^{M}[Pixels(A_i^{k+1})] - Pixels(X(x,y)>0) = 0$, with Pixels and the pixel counting operator. As the growing speed of the previously described algorithm is not dependent on the initial contour and the initial area is given by the thresholded distance transform, this method provides a fast - and due to the initialization - yet precise solution for declustering cells. The described algorithm can be speeded up by varying the size of the structuring element: from iteration to iteration the diameter is decreased. In order to ensure a Minimum of Artifacts, the selection of a suitable initial diameter of the structuring element has to be selected.

3.3. Classification for cell diagnosis using the neural network (NN)

A learning algorithm for back propagation with Kalman filtering is proposed for more efficient training of a neural network for classification. A modified 5-fold cross-validation error estimation technique is proposed, which is a generalization of the leave-one-out technique to achieve reliable system performance and accurate results evaluation, as previously used by our investigators [52]. We run 5-fold cross validation on the data set. In detail, we divide the data set into five subsets with similar size, where the proportion of different classes in each subset is similar to that in the original data set. Then we run each experiment for five times, each time using the union of four subsets as training set to train the lung cancer cell identification module named Kalman filtering Neural Network, and using the remaining subset as test set to test the trained module to see how well it works.

4. RESULTS

4.1. Experimental results from computer-aided cytological diagnosis (CACD) paradigm

As we described in Section 2, the cell lines, immunohistochemistry, and corresponding cell computational features are used and implemented. The different module algorithms in CACD system are applied to the examples of these cellular images, shown in Figure 4 and Figure 5.

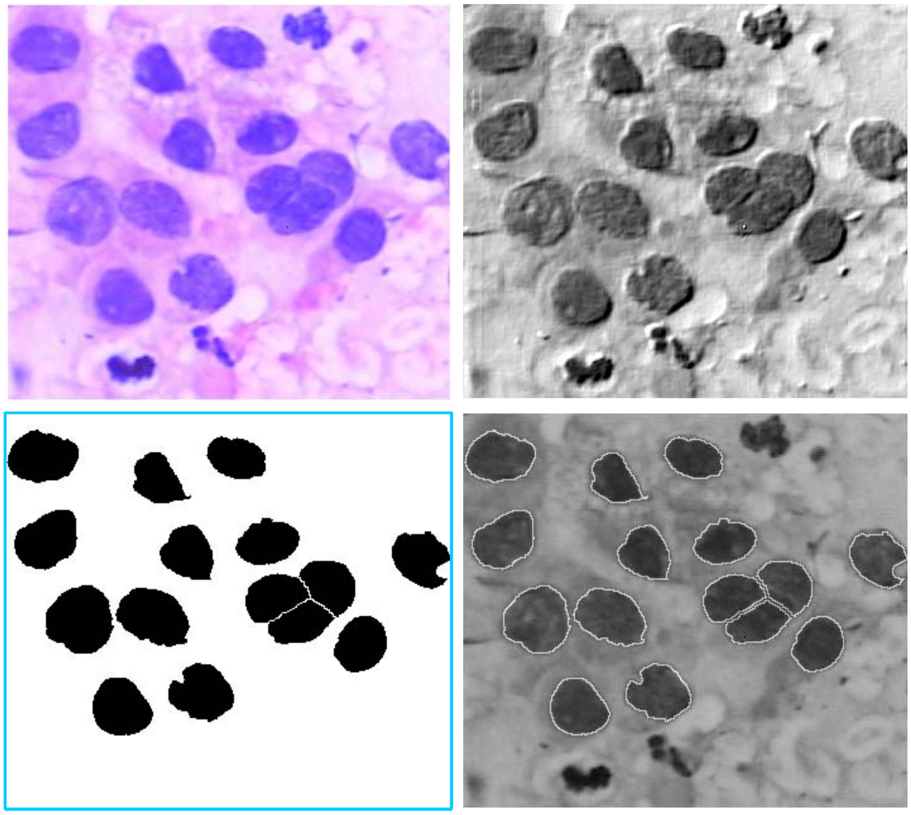


Figure 4

Top left:
Original squamous carcinoma cancer cell image.

Top right:
Enhancement result by using TSF, DWT, WT enhancement modules in CACD

Bottom left:
Pre-segmented result by using proposed Hopfield Neural Network described in Section 3.3

Bottom right:
Result after declustering process by using the algorithm described in Section 3.2

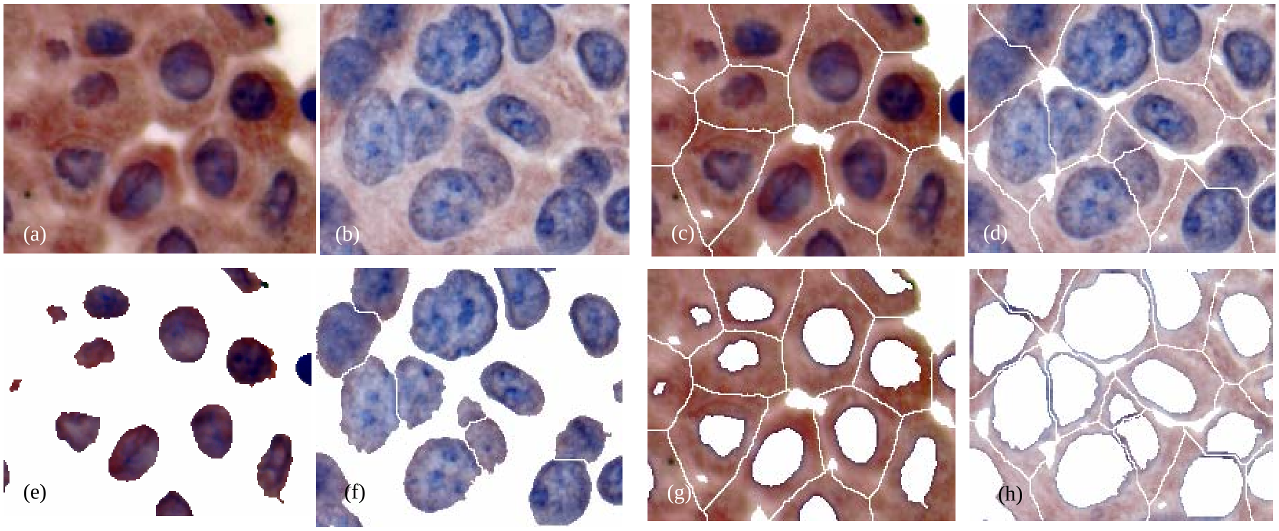


Figure 5. (a).Original normal cell image, (b). Original cancer cell image, (c). Cell segmentation result of normal cell image, (d) Cell segmentation result of cancer cell image. (e). Segmented nuclei of normal cell image. (f) Segmented nuclei of cancer cell image. (g) Segmented cytoplasm tissues of normal cell image. (h) Segmented cytoplasm tissues of cancer image. Microstructure features can be extracted from these images.

4.2. Experiment results on Biomarkers

With bright field microscopy we observed predominately nuclear staining in Squamous cell carcinoma, and nuclear plus cytoplasmic staining in Adenocarcinoma. Of great interest is the strong nuclear antigen expression in cells from an AAH lesion (potentially pre-malignant lesion of the lung) with no detectable γ-H2AX staining in morphologically normal adjacent tissue. Images were obtained with the assistance of the Moffitt Analytic Microscopy Core Facility with the use of Leica DMLB light microscope,

For the initial assessment of biomarker expression with the use of standard immunohisto/cyto-chemistry, we applied cell labeling techniques with visible chromogenes and bright field microscopy, where we can compare different image analysis systems including routinely used in the Analytic Microscopy Core Facility at the Moffitt Research Institute and novel CACD system described in this paper. The software used in the core facility is Spot Advanced (Diagnostic Instruments) ImagePro Plus, version 5.0 to measure color intensities, where the Dynamic Range Density function utilizes to measure the color intensity (units for Dynamic Range run from 255-white to 0-Black). This function allowed to quantitate the mean, min, and max dynamic range. To illustrate the ability of the CACD algorithm to provide mathematical means of histo/cyto-tags for biomarker assessment, we analyzed first images of lung cancer tissue sections immunostained for γ-H2AX. We demonstrated that the CACD algorithm could be applied to extract mean marker expression values for separate features (Figure 6, A-D). Results indicate an elevated expression of γ-H2AX in lung cancer cells when compared to adjacent normal tissue.

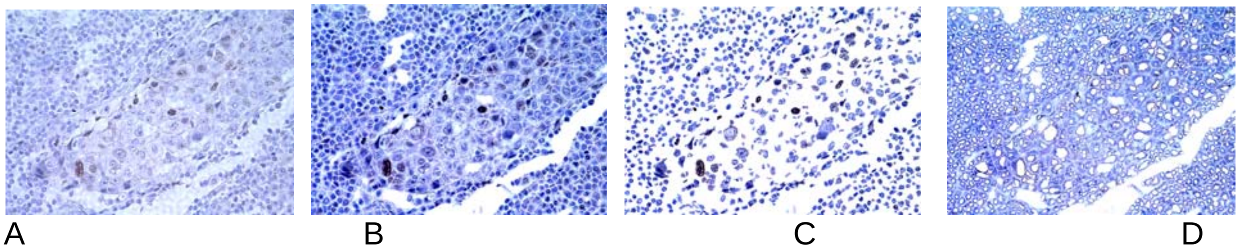


Fig.6A. Raw image of lung tissue- Sq.cell Ca, **Fig.6C**.Enhanced and segmented nuclei image.

γ-H2AX, PX-DAB, 40X. **Fig.6D** Enhanced and segmented cytoplasm image

Fig.6B Enhanced and segmented cell image

In another serial experiments we evaluated ability of the CACD algorithm to provide mathematical means for biomarker assessment (γ-H2AX) on images of immunostained cytospin preparations of cultured non-small cell lung adenocarcinoma cell line A549, Figure 8, shown expression at different dilution of γ-H2AX rabbit antibody (Upstate Biotechnology, Lake Placid, NY). PX-DAB, 40X.

Our data presented in Table III show ability of CACD for the multiple cell feature extraction. There are twelve features computed by our CACD system in Table III. These features are sufficient on classification of cancer versus normal tissues, especially ability to separate signals from nuclei and cytoplasm and prove sub-cellular localization of the targeted cancer-specific biomarkers.

Feature	Area	Convex Area	Equiv. Diam.	Convex Perim.	Skelet. Length	Mean Red	Mean Green	Mean Blue	Mean Hue	Mean Satur	Mean Lumi	Inten. Stdv
2005 pH2AX 40x_Nuclei												
Average	0.118	0.131	0.362	1.29	0.561	139	119	124	190	26.6	127	19
STDV	0.096	0.106	0.14	0.492	0.315	12.7	10.9	15.2	153	7.64	10.5	5.01
Adeno AAH 2005 pH2AX 40x_Nuclei												
Average	0.098	0.107	0.337	1.19	0.47	61.8	44.1	83.8	275	84.6	63.2	22.9
STDV	0.067	0.074	0.107	0.383	0.211	19.5	15.4	23.3	27.7	29	16.1	6.27
Adeno NBrochEpi 2005 pH2AX_Nuclei												
Average	0.055	0.062	0.256	0.928	0.374	126	113	129	211	29.6	123	17.6
STDV	0.033	0.038	0.069	0.253	0.172	20	15.1	18.7	120	10.1	14.4	4.37

Table III. The results are based on three different cell images: normal (Adeno NBrochEpi 2005 pH2AX_Nuclei), pre-malignant lesion (Adeno AAH 2005 pH2AX 40x_Nuclei), and cancer (Adeno 2005 pH2AX 40x_Nuclei). The results are the test of the feasibility of using our presented nuclear features for differentiating the three kinds of cells. From this table, we can see that all the selected features are differentiable (there are distinct differences among the feature measurements of different kind of cells). This test shows a promising prospect that our presented features could be used to identify different kind cells.

Key Research Accomplishments

1. Development of anatomical knowledge structure (AK) modules for AK-CAD

a. Build an AK model about lung structure with nodules using a frame-based semantic network, and incorporate anatomical variability into image processing algorithms using fuzzy sets.

b. Incorporate AK to increase the ability of the CAD system to discriminate anatomical structures, especially those with similar attenuation, which can help increase sensitivity and specificity in detection of lung nodules.

c. Implement an information sharing and coordinating architecture model for communication between the AK model and image processing CAD modules.

2. Generation of helical CT image databases and evaluation of AK-CAD

a. Develop training database containing 200 cases including biopsy proven abnormal cases and normal cases for the system optimization.

b. Construct testing database with another 400 cases containing various types of abnormal cases and normal cases for evaluation of final AK-CAD system.

c. Compare the performance of AK-CAD system with that of CAD system in the detection of lung nodules less than 10 mm in size.

d. Evaluate the overall performance of the AK-CAD system.

3. **Generation of cellular image database and evaluation of CAPD**

a. Based on the detected nodule areas from CT images, cytopathological images are obtained from fine-needle aspiration biopsy (FNAB), bronchial washings and brushing specimens, and histopathological images are obtained from open biopsies or surgical resections. The cellular image database is generated for training and testing CAPD.

b. The cellular image database will comprise 552 representative images. About 75% of those 552 images would represent lung cancer cells with a morphological distribution as follows: 32% adenocarcinoma, 38% squamous cell carcinoma, 22% small cell carcinoma, and about 8% large cell carcinoma, , which represents in most cases a variant of adenocarcinoma. The remaining 25% cases will comprise of benign conditions and will encompass reactive etiologies.

c. These cellular image cases would then be stratified and entered in a common Microsoft Access database to ensure retrieval at a later point of time. Each case will then be reviewed to identify a single representative glass slide. Areas will be identified on the slides under microscope and three representative images will be obtained for each case.

d. Evaluate the overall performance of the CAPD system.

Reportable Outcomes

LIST of JOURNAL PUBLICATIONS

[1]. Qian Wei, Zhukov TA, Song, M.Sc and Tockman MS, "Computerized Analysis of Cellular Features and Biomarkers for Cytologic Diagnosis of Early Lung Cancer." Journal of Analytical and Quantitative Cytology and Histology, Vol. 10, Oct. 2006, PP 18-28.

[2]. Wei Qian, Hong Su and Ravi Sankar "A Knowledge-Based Lung Nodule Detection System for Helical CT Images" International Journal of Computational Intelligence and Applications (IJCIA), will be published in December of 2006.

[3]. Zhukov TA, Pottackal S, Monteiro A, Martino M, Lancaster J, Qian Wei, Song DS, Cantor AB, Sellers TA, Tockman MS. "Novel Lung Cancer Biomarker γ-H2AX (Marker of DNA Damage Response) and Estrogen Receptors Pathways: Do They Crosstalk?" Journal of Thoracic Oncology, In press, 2006

[4]. Luo P, Eikman E, Kealy W, & Qian, W. "An Analysis of a Mammography Teaching Program Based on an Affordance Design Model". Academic Radiology. 12(11): 1112-1123, 2006

[5]. Y.Zhang, R.Sankar and W.Qian, "Boundary Delineation in Transrectal Ultrasound Image for Prostate Cancer", Computers in Biology and Medicine, Octorber 2006 (Submitted).

[6]. Daniel W. McKee, Walker H. Land, Jr., Tatyana Zhukov[c], Dansheng Song, Wei Qian, “An adaptive image segmentation process for the classification of lung biopsy images”, IEEE Trans. on Biomedical Engineering (Submitted in 2006).

[7]. Qian Wei, Minshan Lei, Xiaoshan Song, Ravi Sankar, and Edward Eikman, “Computer Aided Mass Detection based on Ipsilateral Multi-view Mammograms”. Academic Radiology, Accepted for publication in Dec. of 2006.

[8]. Wei Qian, Ph.D, Dansheng Song, MSc. Minsha Lei, MSc. and Edward Eikman, MD “A New Multi-view Computer Aided Detection (CAD) for Multi-view Breast Imaging”, IEEE Signal Processing Magazine, submitted in Nov. of 2006

[9]. Ravi Samala, Moreno Wilfrido, James Leffew, Wei Qian, “A Novel Cancer Region Segmentation for Multi-view Computer Aided Detection for Tomosynthesis Imaging System” IEEE Signal Processing Magazine, submitted in Dec. of 2006

[10]. Minshan Lei, dansheng song, Yudong Yao, Ravi Sankar, Wei Qian, “Shape Similarity and Rotation Invariant Enhancements Using Fourier Descriptors in Digital Mammography", Medical Physics, submitted in Oct. of 2006

[11]. Wei Qian, Ph.D, Tatyana Zhukov, Ph.D, Dansheng Song, MS Nazeel Ahmad, MD, “Adaptive Algorithms for Early Prostate Cancer Cytological Diagnosis**”,** Journal of Physics in Medicine and Biology, submitted in Nov. of 2006

[12]. Luo, P, Eikman, E, Kealy, W, & Qian, W. (2007). “Active E-learning in Mammography Interpretation”. Submitted to Academic Medicine in Dec. of 2006.

[13]. Walker H. Land, Jr.[a] , Richard Lee[b] ,Tatyana Zhukov[c], Dansheng Song[c], Wei Qian[c] “An improved kernel-based adaptive image segmentation process for lung cancer detection from biopsy images” Submitted to IEEE Transactions on Biomedical Engineering in Dec. of 2006.

Conclusion

In PART I: A knowledge-based system for lung nodule detection on helical CT images that was developed has been described. Within modular system architecture, an explicit anatomical model is matched to image data by a blackboard system. The knowledge-based system augments low-level segmentation techniques by allowing high-level interpretation. In this system, domain knowledge provides guidance for object recognition. Using the hierarchy implied by relationships in the model, the blackboard system automatically schedules the identification of anatomical structures. Both ‘a priori’ and ‘posteriori’ information are combined to constrain segmentation of expected anatomy. Fuzzy based classification is used to provide an intuitive representation that allows symbolic description of image feature values, so the high-level rules can be used to generate reports on suspected abnormalities. The preliminary experimental results were encouraging demonstrating the proof of concept of computer aided detection of lung nodule detection using knowledge-based methodology but further validation is required using a more complex model and larger data set.

In PART II: The presented algorithms, CACD system for cellular feature enhancement, segmentation and classification, are very important in distinguishing the benign lesion with malignant lesion. The clearly benign lesions usually have smooth nuclear surface and homogeneous chromatin staining intensity. In contrast, carcinomas displayed remarkably different features in morphology, including: irregular nuclear surface; marked nuclear pleomorphism (irregular, angulated and indented shape of nuclear volume); irregular and coarse chromatin texture and chaotic arrangement of tumor cell nuclei. In conclusion, nuclear structure with morphologic image analysis by using CACD system may provide a useful research diagnostic tool in cytology. And in clinical practice, the dense hyperchromatic cell groups are considered common diagnostic problems in cytopathological evaluations. Cytological evaluations of the dense hyperchromatic groups in cervicovaginal smear results in high rates of false-positive or false negative diagnosis. The key element is to automatically differentiate among the dense hyperchromatic groups and to appropriately classify, based on strict morphologic criteria. Obviously, our proposed CACD system is the good tool.

The automatic extraction of the cancerous nuclei and assess biomarker expression on sub-cellular level in lung pathological color images can segment the images based on chromaticity features and histogram analysis of the RGB color space components of the raw image. The ideal cytomorphometric analysis should differentiate between the ambiguous or suspicious groups of dense hyperchromatic cells. Ultimately, this diagnostic tool, CACD system, can minimize the rate of false-positive or false-negative diagnosis resulting in better cytology/pathology evaluations and patient management.

References

[1] American Cancer Society, "Cancer Statistics 2005 Presentation," http://www.cancer.org/docroot/PRO/content/PRO_1_1_Cancer_Statistics_2005_Presentation.asp, 2005.

[2] L. Ries et al., "SEER Cancer Statistics Review, 1975-2000," National Cancer Institute, Bethesda, MD, 2003.

[3] S. H. Landis et al., "Cancer Statistics, 1999," CA Cancer J. Clin., Vol. 49, No. 1, pp. 8-31, Jan-Feb 1999.

[4] M. Garmer et al., "Digital Radiography Versus Conventional Radiography in Chest Imaging: Diagnostic Performance of a Large-Area Silicon Flat Panel Detector in a Clinical CT-Controlled Study," American J. of Roentgenology, Vol. 174, pp. 75-80, 2000.

[5] C. V. Zwirewich, J. R. Mayo, and N. L. Muller, "Low-Dose High-Resolution CT of Lung Parenchyma," Radiology, Vol. 180, pp. 413-417, 1991.

[6] B. Van Ginneken, B. M. Ter Haar Romeny, and M. A. Viergever, "Computer-Aided Diagnosis in Chest Radiography: A Survey," IEEE Trans. on Medical Imaging, Vol. 20, No. 12, pp. 1228-1241, Dec. 2001.

[7] R. Wiemker, P. Rogalla, and A. Awartkruis, "Computer-Aided Lung Nodule Detection on High Resolution CT Data, Proc. SPIE : Medical Imaging, , Vol. 4684, pp. 677-688, 2002.

[8] B. Van Ginneken et al., “Automatic Detection of Abnormalities in Chest Radiographs Using Local Texture Analysis,” IEEE Trans. on Medical Imaging, Vol. 21, No. 2, pp. 139-149, 2002.

[9] M. S. Brown et al., “Method for Segmenting Chest CT Image Data Using an Anatomical Model: Preliminary Results,” IEEE Trans. on Medical Imaging, Vol. 16, pp. 828-840, 1997.

[10] M. S. Brown et al., “Patient-Specific Models for Lung Nodule Detection and Surveillance in CT Images,” IEEE Trans. on Medical Imaging, Vol. 20, No. 12, pp. 1242-1250, Dec. 2001.

[11] V. Jagannathan, R. Dodhiawala, and L. S. Baum, Blackboard Architectures and Applications, London: Academic Press, 1989.

[12] B. H. Thompspn et al., “Virtual Hospital-Lung Anatomy,” The University of Iowa, http://www.vh.org, 1992.

[13] J. Wang, M. Betke, and J. P. Ko, “Segmentation of Pulmonary Fissures on Diagnostic CT – Preliminary Experience,” International Conference on Diagnostic Imaging and Analysis(ICDIA'02), Shanghai, China, 2002.

[14] G. Sagerer and H. Niemann, Semantic Network for Understanding Scenes, NY: Plenum Press, 1997.

[15] K. Tanaka, An Introduction to Fuzzy Logic for Practical Applications, NY: Springer-Verlag, 1997.

[16] K.Kanazawa et al., “Computer-Aided Diagnosis for Pulmonary Nodules Based on Helical CT Images,” Computerized Medical Imaging and Graphics, pp. 157-167, 1998.

[17] T -H. Cho et a., “A Computer Vision System for Automated Grading of Rough Hardwood Lumber Using a Knowledge-Based Approach,” Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics, pp. 345-350, 1990.

[18] H. Su, W. Qian, R. Sankar, and X. Sun, A New Knowledge-Based Lung Nodule Detection System, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada, May 2004, vol. V, pp. 445-448.

[19] R. Clouard et al., “Borg: A Knowledge-Based System for Automatic Generation of Image Processing Programs,” IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 21, No. 2, pp. 128-144, Feb. 1999.

[20] B. Sahiner et al., “Image Feature Selection by a Genetic Algorithm: Application to Classification of Mass and Normal Breast Tissue on Mammograms”, Med. Phys., Vol. 23, pp.1671-1648, 1996.

[21] K. Kanazawa et al., “Computer-Aided Diagnosis for Pulmonary Nodules Based on Helical CT Images,” Computerized Medical Imaging and Graphics, Vol. 22, pp. 157-167, 1998.

[22] S. Chuai-Aree, C. Lursinsap, P. Sophasathit, and S. Siripant, “Fuzzy C-Mean: A Statistical Feature Classification of Text and Image Segmentation Method,” Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol.9, No.6, pp. 661-671, 2001.

[23] M. N. Ahmed et al., “A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation of MRI Data,” IEEE Trans, on Medical Imaging, Vol. 21, No. 3, pp. 193-199, Mar. 2002.

[24] N. Petrick et al., “An Adaptive Density-weighted Contrast Enhancement Filter for Mammographic Breast Mass Detection,” IEEE Trans. on Medical Imaging, Vol.15, No.1, pp. 59-67, Feb. 1996.

[25]. Dillon, D.A., et al. "The Molecular Biology of Breast Cancer: Accelerating Clinical pplications." Critical Reviews in Oncogenesis 9.2 (1998):125-140.

[26]. Burma,S., Chen,B.P., Murphy,M., Kurimasa,A. & Chen,D.J. ATM phosphorylates histone H2AX in response to DNA double-strand breaks. J. Biol. Chem, 276: 42462-42467, 2001.

[27]. Ward,I.M. & Chen,J. Histone H2AX is phosphorylated in an ATR-dependent manner in response to replicational stress. J. Biol. Chem. 276, 47759-47762, 2001.

[28]. Bassing,C.H. et al. Histone H2AX: a dosage-dependent suppressor of oncogenic translocations and tumors. Cell 114, 359-370, 2003.

[29]. Sedelnikova OA, Pitch DR, Redon C, Bonner WM. Histone H2AX in DNA damage and repair. Cancer Biol Ther 2(3):233-235, 2003.

[30] Garrido A, Perez N. Applying deformable templates for cell image segmentation. Pattern Recognit 2000; 33:821–32.

[31] Mouroutis T, Roberts SJ, Bharath AA. Robust cell nuclei segmentation using statistical modeling. BioImaging 1998; 6:79–91.

[32] Simon I, Pound CR, Partin AW, Clemens JQ, Christensbarry WA. Automated image analysis system for detecting boundaries of live prostate cancer cells. Cytometry 1998; 31:287–94.

[33] Wu HS, Barba J, Gil J. A parametric fitting algorithm for segmentation of cell images. IEEE Trans Biomed Eng 1998; 45:400–7.

[34] Wu HS, Barba J, Gil J. Iterative thresholding for segmentation of cells from noisy images. J Microsc 2000; 197:296–304.

[35] Kapur JN, Sahoo PK, Wong AKC. A new method for gray-level picture thresholding using the entropy of the histogram. Comput Vis Graph Image Process 1985; 29:273–85.

[36] Kittler J, Illingworth J. Minimum error thresholding. Pattern Recognit 1986;19:41–.

[37] Hamasaki M, Kamma H, Wu W, and others. Expression of hnRNP B1 in four major histological types of lung cancers. Anticancer Res. 2001; 21(2A): 979-84

[38] Theocharis S, Kanelli H, Politi E, and others. Expression of peroxisome proliferator activated receptor-gamma in non-small cell lung carcinoma: correlation with histological type and grade. Lung Cancer (Lung cancer (Amsterdam, Netherlands)) 2002; 36(3): 249-55

[39] Zhou ZH; Jiang Y; Yang YB, and others. Lung cancer cell identification based on artificial neural network ensembles. Artif Intell Med (Artificial intelligence in medicine.) 2002; 24(1): 25-36

[40] Yamaji H; Iizasa T; Koh E, and others. Source Correlation between interleukin 6 production and tumor proliferation in non-small cell lung cancer. Cancer Immunol Immunother (Cancer immunology, immunotherapy : CII.) 2004; 53(9): 786-92.

[41] Traynor AM; Schiller JH. Systemic treatment of advanced non-small cell lung cancer. Drugs Today (Barc) (Drugs of today (Barcelona, Spain : 1998)). 2004; 40(8): 697-710.

[42] Tanno S; Ohsaki Y; Nakanishi K, and others. Small cell lung cancer cells express EGFR and tyrosine phosphorylation of EGFR is inhibited by gefitinib ("Iressa", ZD1839). Oncol Rep (Oncology reports.). 2004; 12(5): 1053-7.

[43] Bergqvist M; Brattström D; Larsson A, and others. The role of circulating anti-p53 antibodies in patients with advanced non-small cell lung cancer and their correlation to clinical parameters and survival. BMC Cancer (BMC cancer [electronic resource].) 2004; 4(1): 66.

[44] Campione A; Ligabue T; Luzzi L, and others. Impact of size, histology, and gender on stage IA non-small cell lung cancer. Asian Cardiovasc Thorac Ann (Asian cardiovascular & thoracic annals.) 2004; 12(2): 149-53

[45] Mao, F., Qian, W., Gaviria ,J. and Clarke, LP. "Fragmentary Window Filtering for Multiscale Lung Nodule Detection: Preliminary Study," *Academic Radiology*, Vol. 5, No.4, April 1998, pp306-311.

[46] W. Qian, L. P. Clarke, M. Kallergi, R. A. Clark, "Tree-structured nonlinear filters in digital mammography," IEEE Trans. Med. Imag., vol. 13(1), 25-36, 1994.

[47] L. Li, W. Qian and L.P. Clarke, "Digital mammography: CAD method for mass detection using multiresolution and multiorientation wavelet. transforms," Academic Radiology, 1997; 4:724-731.

[48] W. Qian, L.P. Clarke, L. Li, et. al., "Computer Assisted Diagnostic (CAD) Methods for X-ray Imaging and Teleradiology," Proceedings of the 26th AIPR Workshop, Cosmos Club, D.C., Oct. 15-17, 1997.

[49] W. Qian, M. Kallergi and L. P. Clarke" Order Statistic-Neural Network Hybrid Filters for Gamma Camera Image Restoration. " IEEE Trans. in Medical Imaging, pp56-65, March, 1993.

[50] W. Qian, L. P. Clarke, "Wavelet-based neural network with fuzzy-logic adaptivity for nuclear image restoration," Proceedings of the IEEE, Special Issue on Applications of Neural Networks, Invited paper, vol.84, no. 10, 1996.

[51] R. Sammouda, N. Niki and H. Nishitani , A comparison of Hopfield Neural Network and Boltzmann machine in segmenting MR images of the brain. *IEEE Trans. Nucl. Sci.* 43 6 (1996).

[52] B. Y. Zheng, W. Qian, L. P. Clarke, " Digital mammography: MF-based NN for automatic detection of microcalcifications," IEEE Trans. on Medical Imaging, pp.589-597, Oct. 1996.

Appendices

Three papers in these Appendices are in preparation for publication.

1. Aleksandra Zajac, Dansheng Song, Wei Qian and Tatyana Zhukov, “Protein Microarrays and Quantum Dot-Probes for Early Cancer Detection” To be submitted to Journal of Thoracic Oncology.

2. Tatyana Zhukov*, Sangy Pottackal, Aleksandra Zajac, Rachna Kapoor,
Dan Sheng Song, Wei Qian, Ambuj Kumar, Melvyn Tockman, “ Does high expression of estrogen receptor and DNA damage signaling (γ-H2AX) indicate poor prognosis in female patients with resected Nonsmall Cell Lung Cancer? “ To be submitted to Journal of Thoracic Oncology.

3. Daniel W. McKeea, Walker H. Land, Jr.b, Tatyana Zhukovc, Dansheng Songc, Wei Qian, “An adaptive image segmentation process for the classification of lung biopsy images” To be submitted to Journal of Medical Physics

Protein Microarrays and Quantum Dot-Probes for Early Cancer Detection

Aleksandra Zajac, Dansheng Song, Wei Qian and Tatyana Zhukov*

Division of Cancer Prevention and Control – H. Lee Moffitt Cancer Center and Research Institute at the University of South Florida, 12902 Magnolia Drive, Tampa, FL 33612, USA

*Corresponding author: Dr Tatyana Zhukov, PhD, PMIAC Department of Interdisciplinary Oncology, H. Lee Moffitt Cancer Center and Research Institute, University of South Florida, 12902 Magnolia Drive, Tampa, FL 33612, USA.

Tel: 813 745 1718; fax: 813 745 6525

E-mail: zhukovta@moffitt.usf.edu

ABSTRACT

We describe here a novel approach for detection of cancer markers using Quantum Dot Protein Microarrays. Both relatively new technologies; quantum dots and protein microarrays, offer very unique features that together allow detection of cancer markers in biological specimens (serum, plasma, body fluids) at pg/ml concentration. Quantum dots offer remarkable photostability and brightness. They do not exhibit photobleaching common to organic fluorophores. Moreover, the high emission amplitude for QDs results in a marked improvement in the signal to noise ratio of the final image. Protein microarrays allow highly parallel quantitation of specific proteins in a rapid, low-cost and low sample volume format. Furthermore the multiplexed assay enables detection of many proteins at once in one sample, making it a powerful tool for biomarker analysis and early cancer diagnostics.

In a series of multiplexing experiments we investigated ability of the platform to detect six different cytokines in protein solution. We were able to detect TNF-α, IL-8, IL-6, MIP-1β, IL-13 and IL-1β down to femtomolar concentration, demonstrating high sensitivity of the investigated detection system.

We have also constructed and investigated two different models of quantum dot probes. One by conjugation of nanocrystals to antibody specific to the selected marker – IL10, and the second by use of streptavidin coated quantum dots and biotinylated detector antibody. Comparison of those two models showed better performance of streptavidin QD – biotinylated detector antibody model. Data quantitated using custom designed computer program (CDAS) show that proposed methodology allows monitoring of changes in biomarker concentration in physiological range.

Keywords: quantum dots, protein microarrays, biomarkers, cytokines, cancer detection

INTRODUCTION

According to American Cancer Society, in 2006 there will be over 1.3 million cancer cases in United States and over 500 000 of them will result in death [1]. Although there is an ongoing research in many medical fields to improve patients' outcome, early detection seems to be the key for cancer survival.

It is believed that availability of multiple biomarkers is extremely important in the diagnosis of complex diseases like cancer [2,3], and tests for single markers like for example CA125 for ovarian cancer are not adequate. Unfortunately, only 80% of patients have elevated levels of CA125 [4]. The situation is even worse for patients with early stage disease, where less than 50% of cases have elevated levels [5]. Patterns of multiple cancer markers might provide sufficient information of disease diagnosis in its early stages [3]. However, ability to use this approach in disease diagnosis is dependent on the technology that will allow healthcare professionals for multiplexed and fast detection of many biomarkers simultaneously with high sensitivity and specificity. We believe that proposed platform of quantum dot protein microarrays offers unique features making it possible to fulfill the task.

Quantum Dots are semiconductor nanocrystals structures widely used in bioimaging applications. They can be made from a variety of inorganic compounds dependent upon their mode of application. QDs usually consist of< 10-nm CdSe semiconductor core surrounded by an inorganic shell composed of ZnS. The core-shell complex is coated with a polymer to make the particle water soluble, followed by functionalization with, for example, streptavidin to prepare QDs for use in immunochemistry. The final size of immunochemically functional QDs is c. 10-15 nm, which is in size range of macromolecules [6,7].

QDs are characterized by broad absorption band and narrow symmetric emission band. Importantly for use as biological probes, QDs absorb and emit light through wide spectrum of wavelengths, from visible to NIR [8], and can be excited with any wavelength from UV to red [9,10]. In comparison to organic fluorophores, QDs absorb excitation photons in wider spectral range and emit photons more efficiently due to their higher quantum yield which is an extremely important feature for sensitive fluorescence imaging [8,10-12]. Quantum Dots are resistant to photobleaching and approximately thousand times more photostable than organic dyes. They also exhibit much longer life time [10,13-15] than organic dyes. High signal to noise ratio results in good contrast images, allowing easy separation of QDs from background fluorescence.

Quantum Dots have large Stokes shift values (~ 300-400nm, depending on the excitation wavelength), which leads to improved sensitivity of detection [10,16].

Ability of being excited by single source makes QDs really unique label for multiplexed experiment, where different species of QDs can simultaneously track various biomarkers and unlike organic dyes, be excited with the same source [17]. Bioconjugated Quantum Dot probes have demonstrated their potential to be useful for cancer diagnosis through diverse approaches, like e.g: in vitro diagnostic assays (protein biomarker detection, nucleic acid biomarker detection, high-throughput multiplexing); cellular labeling (fixed cells and tissues, live cell imaging) and in vivo imaging (vascular imaging, lymph node tracking, tumor targeting and imaging) [8].

Protein Microarrays. Mark Schena - author of the first paper demonstrating usefulness of microarrays [18], describes them as analytical devices that possess four distinct characteristics: (a) microscopic target elements or spots, (b) planar substrates, (c) rows and columns of elements and (d) specific binding between microarray target elements on the substrate and probe molecules in the solution [19]. It is a miniaturized assay where each spot contains “bait” molecules (antibody in protein / antibody microarrays), which are probed with unknown biologic sample containing analytes of interest [20]. By processing the microarrays with detector antibody tagged with fluorescent label, each spot produces fluorescent signal proportional to the analyte of interest present in the solution and captured/bound to the “bait” molecule [21].

Applications of functional protein microarrays include: expression profiling for identification and quantitation of proteins present in the solution; protein-protein interactions for examination of binding activity and binding partners of proteins across entire proteome; drugs for identification of drugs activity, targets, cross-reactivity, and diagnostics to measure proteins expressed in serum samples [19].

Production of microarrays consists of the following steps: printing and immobilization of capture antibodies on a functionalized surface (usually glass slide covered with poly-L-lysine, aldehyde, epoxy or nitrocellulose) [21]; incubation with the sample, detection with fluorescent probe, image capture and analysis. The most sensitive method for protein microarrays processing is the “sandwich assay” based on Elisa technique. It utilizes two antibodies that simultaneously bind to the same antigen: “capture” immobilized onto the surface and “detector” fluorescently labeled, producing a fluorescent signal.

Intensity of the fluorescent signal is the representation of biomarker concentration in the solution.

Bridging together unique features of Quantum Dots and protein microarrays can lead to design of very sensitive, robust and feasible assay for early detection of cancer. Massive multiplexing capabilities of Quantum Dots for detection of many cancer biomarkers simultaneously, their exceptional brightness and stability together with biochip approach of protein microarrays, hold tremendous promise for unraveling complex gene expression profiles of cancers and for accurate early clinical diagnosis [8].

MATERIALS AND METHODS

Probe evaluation. To choose the best QD probe, we constructed and evaluated two models. One by conjugation of nanocrystals to antibody specific to the selected marker – IL10, and the second by use of streptavidin coated quantum dots and biotinylated detector antibody. For our evaluation experiments we chose only one QD size (655) however, for multiplexing purposes it is possible to use also other sizes. Monoclonal capture (clone JES3-9D7) rat anti-human

Interleukin-10 antibody, detector (clone JES3-12G8) rat anti-human Interleukin-10 antibody, and recombinant human Interleukin-10 were purchased form Serotec (Raleigh, NC). Microarray nitrocellulose coated glass slides were purchased from Whatman (Sanford, ME) and Quantum Dot 655 antibody conjugation kit was purchased from Invitrogen (Carlsbad, California). The conjugation of detector rat anti-IL10 antibody to QD655 was performed according to the protocol. “Probe evaluation” array was spotted with BioRobotics MicroGrid microarrayer from Genomic Solutions (Ann Arbor, MI) using Stealth printing pins (TeleChem International, Sunnyvale, CA). Six arrays of capture rat anti-IL10 antibody at 0.5mg/ml (1:2 dilution in printing buffer (TeleChem International, Sunnyvale, CA)) were printed on nitrocellulose coated glass slides. After spotting the slides were placed in a box at 4°C overnight. The next morning slides were rinsed with PBS (Sigma-Aldrich, Germany) and blocked (Whatman; Sanford, ME) for one hour. After rinsing again with PBS slides were incubated for 2 h with human IL-10 solutions in PBS at 500 ng/ml (2 arrays), 1000 pg/ml (2 arrays) and 100 pg/ml (2 arrays) and then rinsed again with PBS. Half of the arrays was incubated with detector rat anti-IL10 antibody conjugated to QD655 (20nM) and the remaining half with biotinylated rat anti-IL10 antibody for 1 h with gentle rocking. Following the incubation, slides were rinsed with PBS and three arrays, previously incubated with biotinylated detector antibody, were incubated for 30 min with streptavidin QD655 (Invitrogen; Carlsbad, CA), diluted (1:50, 20nM) with Tris–Buffered Saline (Dako; Carpinteria, CA). Slides were first rinsed with PBS then water and centrifuged dry at 2500 rpm for 3 min. Slides were imaged under fluorescent microscope (Nikon Eclipse E800) equipped with Qdot655 filters and quantitated using custom designed software which utilizes computerized dynamic analysis system (CDAS) for classifying microarrays to measure the features on spots: such as area and intensity [22,23].

CDAS. The analysis system consisted of four modules: A) Edge detection on the image. Canny edge detector was used for detection of spot edges. B) Hough transform computation. After computation of the radii of the spots present in the image the accumulator array corresponding to each of the above radii was filled. Each array was composed of cells for the (x,y) coordinates of the center of the potential circle (boundary of the spot). The default size of the cells was defined to be 4x4 pixels. The hough-circle transformation was applied. For each edge pixel it first run through a sequence of x-values and computed the corresponding y-values for that radius. Then it run through a sequence of y-values and computed the corresponding x-values for that radius. The sequence of x-values varied from x(edge-pixel) -(radius/cos(45)) to x(edge-pixel) + (radius/cos(45)). The same was true for the sequence of y-values. The two sequences were processed in this manner because as the points reached the x (y)-axis, we got the same y(x)-value for different x(y)-values, for points lying on the circle corresponding to the hough transform. This choice of sequences did not let any bias to be introduced because of choice of an x or a y sequence. C) Circle detection. Once the hough transform image for a particular radius was computed, it was adjusted to lie between 0 and 1 and thresholded. In this way only those points with high probability of being the centers were left. The resulting point-sets were labeled with different regions. The centroids of each region were considered as centers of the detected spots. The output image was computed by drawing circles with these points as centers, the radii were matched and the image was segmented. D) Feature extraction. After spots segmentation, the feature extraction was implemented. The average intensity of the spots and standard deviation were calculated.

Multiplexing experiment. The slides for multiplexed detection of TNF-α, IL-8, IL-6, MIP-1β, IL-13 and IL-1β cytokines were purchased from Allied Biotech (Ijamsville, MD). Each array consisted of quadruplicates of capture antibodies against the cytokines, negative and positive controls. A series of eight four-fold step dilutions of cytokines was performed with provided dilution buffer. Slides were washed for 15 min with provided washing buffer and eight concentrations of cytokine standards (range: 4ng/ml – 250 fg/ml, except for IL-13: 16ng/ml – 1pg/ml) were added to the microarrays for 2 h incubation. Slides were washed and a cocktail of biotinylated detector antibodies was added for 1 h incubation. Following the incubation, slides were rinsed with washing buffer and streptavidin-QD655 diluted in Tris–Buffered Saline (1:50) was added to the arrays for 30 min. After washing, the slides were imaged under fluorescent microscope (Nikon Eclipse E800) equipped with Qdot655 filters and quantitated using custom designed software (CDAS).

Data collection. Nikon Eclipse E800 microscope equipped with 2.5x objective and filter sets for QD655 was used for fluorescence imaging. Fixed exposure time was set up for all images for direct comparison of fluorescent data. Images were captured with cooled CCD digital camera (model RTE/CCD-1317-K/2). The net fluorescence intensity (after background removal) was obtained using custom design software (CDAS). The statistical analysis were carried out using Microsoft Excel.

RESULTS

We have divided our investigation into two parts: a) evaluation of two models of QD probes and b) multiplexed detection of six different cytokines: TNF-α, IL-8, IL-6, MIP-1β, IL-13 and IL-1β. In the first part of our experiment we have spotted six 3x6 arrays (130 цm spot diameter) of capture rat anti-IL10 (IgG1) antibody and incubated them with human IL10 at concentrations 500 ng/ml (2 arrays), 1000 pg/ml (2 arrays) and 100 pg/ml (2 arrays). For detection we used QD655 conjugated to rat anti-IL10 (IgG2a) antibody and streptavidin QD655 with biotinylated rat anti-IL10 (IgG2a) antibody. After capture of fluorescent images, we have observed that spots of the array incubated with lowest concentration of IL10 (100 pg/ml) and detected with conjugated QD were not visible, and numerical data analysis of the remaining arrays showed approximately thirty times higher intensity of spots probed with biotinylated anti–IL10 and streptavidin QD655, compared to spots probed with anti-IL10 conjugated to QD655 (Fig. 1).
This result can be attributed to the fact that streptavidin/biotin system, in comparison to QD/antibody conjugate, has much higher energy of association for protein binding in aqueous solution [24,25]. The complexes are also extremely stable over a wide range of temperature and pH.

Having chosen the QD probe type we carried out a series of multiplexing experiment for detection of six cytokines in buffer solution. Nitrocellulose slides spotted with capture antibodies against TNF-α, IL-8, IL-6, MIP-1β, IL-13 and IL-1β were incubated with eight mixes of four-fold step cytokine dilutions in the buffer and detected with biotinylated Ab/streptavidin QD655 complex.
Plots of data obtained through analysis of fluorescent images showed the sensitivity of the assay as low as femtogram range (Fig. 2).

To demonstrate specificity of detection the capture antibody arrays against IL-2, IFN-γ, TNF-α, IL-8, IL-12p70, IL4, IL6, IL-10, IL-12p40, IL-5, IP-10, MIP-1β, IL-13, IL-1β were incubated with 6 different concentrations of IL10 cytokine. The results presented in Figure 3 indicate that detection was specific and no-cross reactivity was observed.

DISCUSSION

In this report we have demonstrated a construction of a sensitive, multiplexed microarrays assay where we incorporated semiconductor Quantum Dots as a label in a sandwich immunoassay for assessing of markers detectable in serum/plasma that potentially offer attractive approaches to develop cancer screening and diagnostic tests.

We applied described system for detection of human cytokines – as an appropriate clinically relevant approach for cancer biomarkers detection. We developed a prototype of a serologic assay with greater sensitivity than current methods and illustrate our method with chosen serologic markers for cancer – panel of cytokines. Cytokines are proteins with the ability to stimulate or inhibit cell growth, regulate cell differentiation, induce cell chemotaxis and modulate the expression of other cytokines. Generally, in tumors there is a shift in the balance of expression of cytokines resulting in enhanced cell proliferation, angiogenesis and metastasis. They can also block immune cell-mediated mechanisms for identifying and destroying tumor cells and may be indicative of a poor prognosis. Pro-inflammatory cytokines such as TNF- α and a number of different interleukins have been identified as orchestrating pre-neoplastic process in a number of cancers [26]. For certain types of cytokines levels were reported increased in sera of cancer patients compared with healthy controls, i.e., IL-1-β, IL-6, IL-8, TNF- α; while for others such as IL-10 level - it has been found tendency to decrease with cancer, that require super-sensitive and specific methods for their detection.

The superior brightness and no-bleaching of QDs demonstrated to be key features for sensitive biomarker detection. We were able to detect simultaneously several human cytokines down to femtomolar concentration whereas commercially available ELISA systems allow detection of cytokines at 5-10 pg/ml concentration range. Multiplexed approach combined with glass slide microarray format proved to be very feasible and efficient method for biomarker detection studies. We have also constructed and evaluated two types of QDs probes where streptavidin QDs/biotinylated detector Ab model showed higher sensitivity than QD/detector Ab conjugate. Our further investigations are going to be concentrated on the use of described system in human biological specimens for assessing cancer relevant circulating and cell-based biomarkers.

ACKNOWLEDGEMENTS

We would like to thank Professor Thomas A. Sellers, Associate Center Director, Cancer Prevention & Control Division at the H. Lee Moffitt Cancer Center & Research Institute for his invaluable help and support, as well as for encouragement and inspiration to pursue pioneering work in the field of nanotechnology and early cancer biomarkers detection.

REFERENCES

[1] American Cancer Society, www.cancer.org. (2006)

[2] J. D. Wulfkuhle,.; L. A. Liotta,.; E. F. Petricoin,. Nat. Rev. Cancer 3 (2003), 267.

[3] G. Zheng, F. Patolsky, Y. Cui, W. U Wang,. C. M. Lieber,. Nat. Biotechnol. 23 (2005), 1294.

[4] J. S. Barnholtz-Sloan; A. G. Schwartz, F. Qureshi, S.Jacques,J. Malone, A. R. Munkarah, Am. J. Obstet. Gynecol. 189 (2003), 1120-.

[5] H. A. Fritsche, R. C. Bast, Clin. Chem. 44 (1998), 1379.

[6] Invitrogen company. www.qdots.com.

[7] W. C. Chan,D. J. Maxwell, X. Gao, R. E. Bailey, M. Han, S. Nie, Curr. Opin. Biotechnol. 13 (2002), 40.

[8] A. M. Smith, S. Dave, S. Nie, L. True, X. Gao, Expert Rev. Mol. Diagn. 6 (2006), 231.

[9] F. Tokumasu, J. Dvorak, J. Microsc. 211 (2003), 256.

[10] S. Santra, D. Dutta, G. A. Walter, B. M. Moudgil, Technol. Cancer Res. Treat. 4 (2005), 593.

[11] B. Dabbousi, J. Rodriquez-Viejo, F. Mikulec, J. Heine, H. Mattoussi, R. Ober, K. Jensen, M. Bawendi, Journal of Physical Chemistry B 101 (1997), 9463.

[12] W. C. Chan, S. Nie, Science 281 (1998), 2016.

[13] C. B. Murray, D. J. Norris, M. G. Bawenti, Journal of American Chemical Society 115 (1993), 8706.

[14] D. Gerion, F. Pinaud, S. C. Williams,; W. J. Parak,; D. Zanchet, S. Weiss, A. Alivisatos, Journal of Physical Chemistry B 105 (2001),.

[15] N. Gaponik, D. V. Talapin, A. L. Rogach, K. Hoppe, E. V. Shevchenko,; A. Kornowski, A. Eychmuller, H. Weller, Journal of Physical Chemistry B 106 (2002), 7177.

[16] X. Gao, Y. Cui, R. M. Levenson, L. W. Chung, S. Nie, Nat. Biotechnol. 22 (2004), 969.

[17] X. Gao, S. Nie, Trends Biotechnol. 21 (2003), 371.

[18] M. Schena, D. Shalon, R.W. Davis, P. O. Brown, Science 270 (1995), 467.

[19] M. Schena, Protein Microarrays, Jones and Bartlett Publishers, Sudbury, 2005, Chapter 1.

[20] L. A. Liotta, V. Espina, A. I. Mehta, V. Calvert, K. Rosenblatt, D. Geho, P. J. Munson, L. Young, J. Wulfkuhle,; E. F. Petricoin, 3rd. Cancer Cell 3 (2003), 317.

[21] W. Kusnezow, A. Jacob, A. Walijew, F. Diehl, J. D Hoheisel,. Proteomics 3 (2003), 254.

[22] L. Qin,; L. Rueda,; A. Ali,; A. Ngom, Appl. Bioinformatics 4 (2005), 1.

[23] X. H. Wang, R. S. Istepanian, Y. H. Song, IEEE Trans. Nanobioscience 2 (2003), 190.

[24] S. Miyamoto, P. A. Kollman, Proteins 16 (1993), 226.

[25] S. Miyamoto, P. A Kollman,. Proc. Natl. Acad. Sci. U S A 90 (1993), 8402.

[26] S. C. Robinson, L. M. Coussens, Adv. Cancer. Res. 93 (2005), 159.

FIGURE CAPTIONS

Fig. 1 Spot intensity comparison for two different QD/antibody probe types.

The bars represent averaged intensity of spots from arrays incubated with human IL10 at 500 ng/ml and 1000 pg/ml concentration. Spots from array incubated with IL10 (100pg/ml) and detected with conjugated QD/antibody complex were not visible.

Fig. 2 Sensitivity of capture. Arrays of: anti - IL-6 (A), anti - MIP-1β (B), anti- IL-13 (C), anti-IL-8 (D), anti-TNF-α (E), anti-IL1β (F) were used to determine the sensitivity of the multiplexing capabilities of the assay. The concentration of cytokines in the solutions for TNF-α, IL-6, IL-8, MIP-1β, IL1β were: 4 ng/ml, 1 ng/ml, 250 pg/ml, 62.5 pg/ml, 16 pg/ml, 4 pg/ml, 1 pg/ml, 250 fg/ml and for IL13 the concentration in the solutions were: 16 ng/ml, 4 ng/ml, 1 ng/ml, 250 pg/ml, 62.5 pg/ml, 16 pg/ml, 4 pg/ml, 1pg,ml.

Fig 3. Specificity of detection. Capture antibodies against IL-2, IFN-γ, TNF-α, IL-8, IL-12p70, IL4, IL6, IL-10, IL-12p40, IL-5, IP-10, MIP-1β, IL-13, IL-1b were incubated with IL-10 at 500 ng/ml, 50 ng/ml, 5ng/ml, 50pg/ml, 5pg/ml, 500fg/ml. The IL-10 antigen was captured only by its cognate antibody.

FIGURES

Figure 1

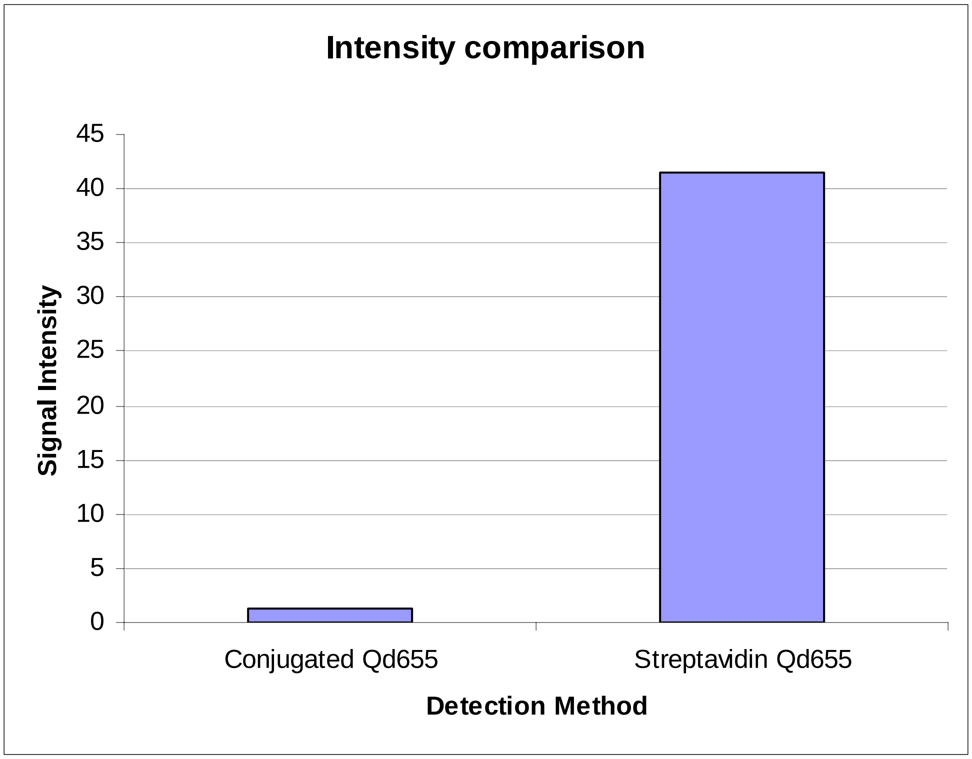
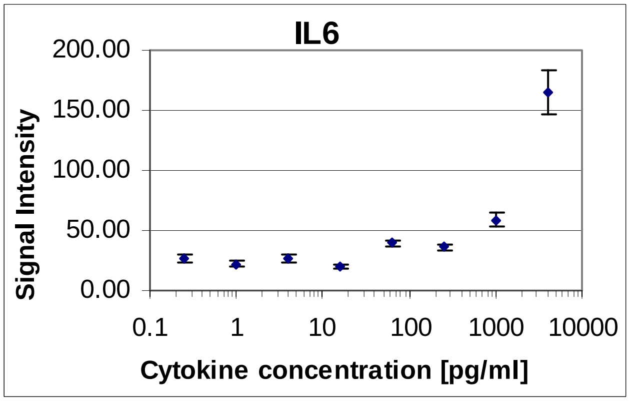
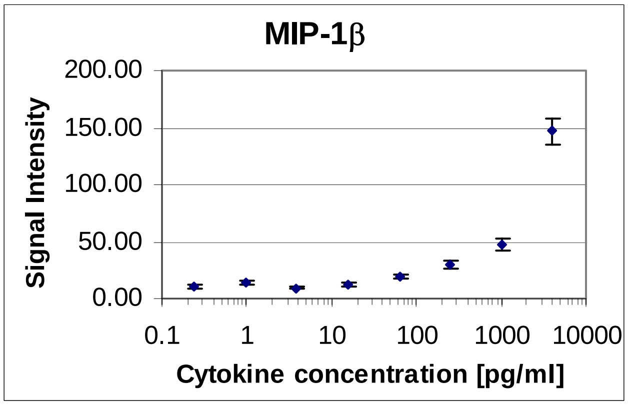


Figure 2

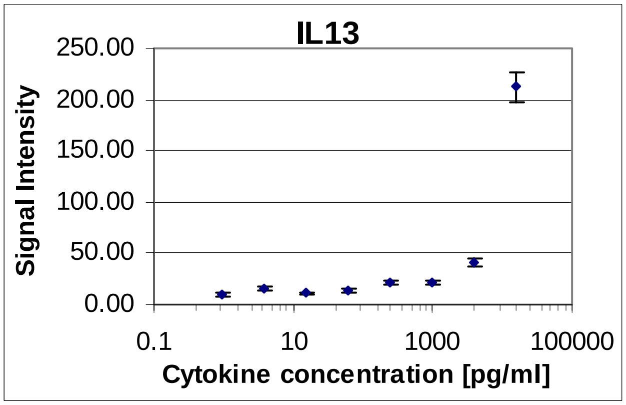
(A)



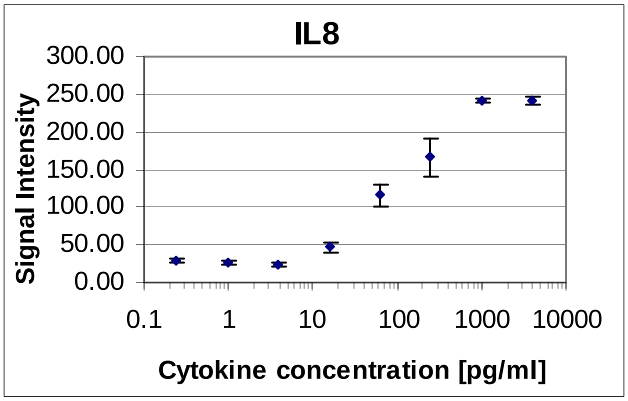
(B)


MIP-1β
Signal Intensity
200.00
150.00
100.00
50.00
0.00
0.1
1
10
100
1000
10000
Cytokine concentration [pg/ml]

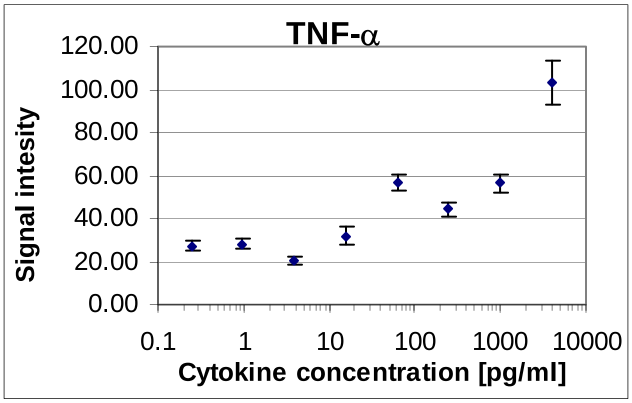
(C)


IL13
Signal Intensity
250.00
200.00
150.00
100.00
50.00
0.00
0.1
10
1000
100000
Cytokine concentration [pg/ml]

(D)


IL8
Signal Intensity
300.00
250.00
200.00
150.00
100.00
50.00
0.00
0.1
1
10
100
1000
10000
Cytokine concentration [pg/ml]

(E)


TNF-α
120.00
100.00
80.00
60.00
40.00
20.00
0.00
Signal intesity
0.1
1
10
100
1000
10000
Cytokine concentration [pg/ml]

(F)

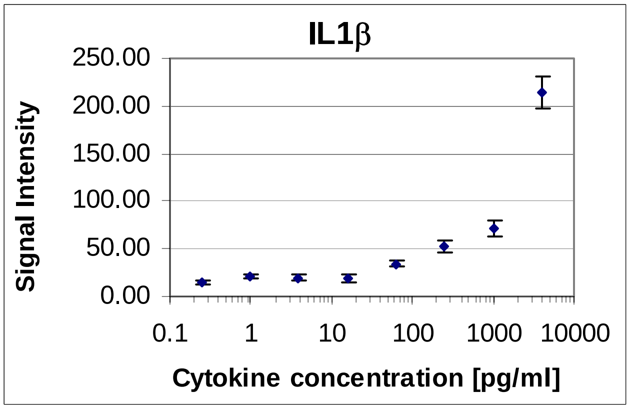
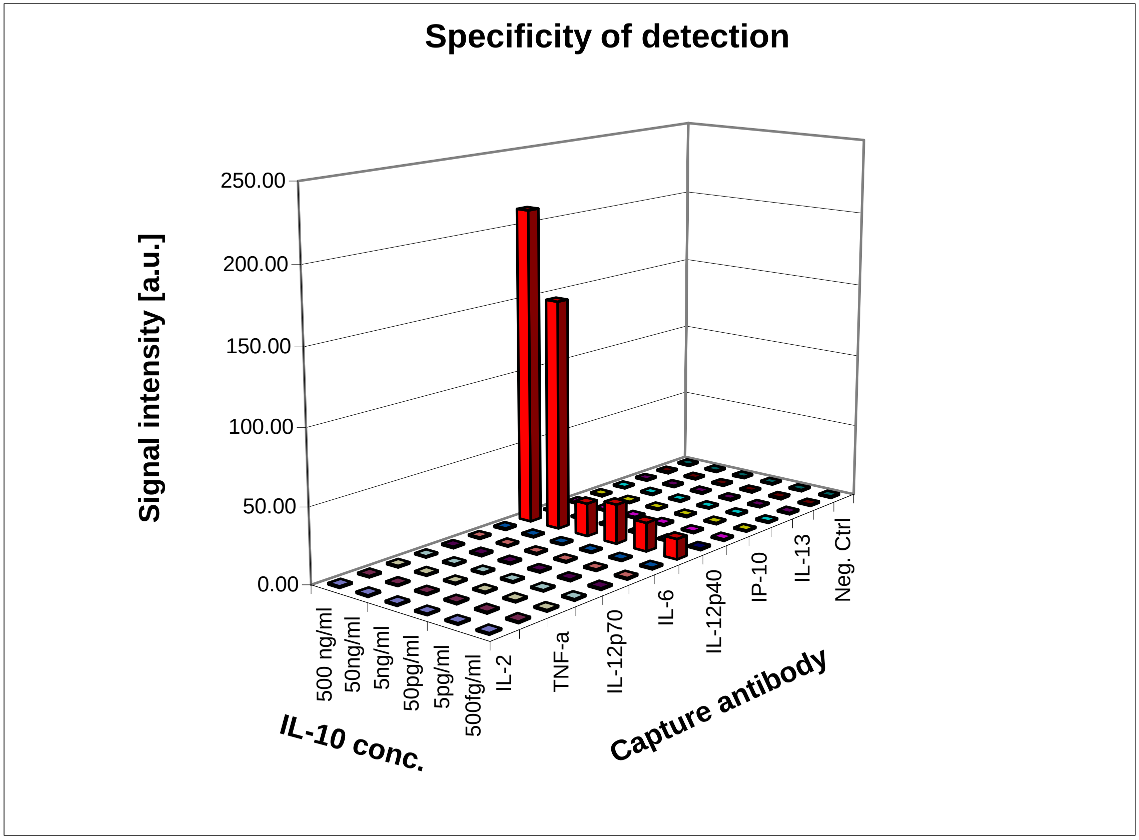

IL1β
250.00
200.00
150.00
100.00
50.00
0.00
Signal Intensity
0.1
1
10
100
1000
10000
Cytokine concentration [pg/ml]

Figure 3


Specificity of detection
Signal intensity [a.u.]
250.00
200.00
150.00
100.00
50.00
0.00
500 ng/ml
50ng/ml
5ng/ml
50pg/ml
5pg/ml
500fg/ml
IL-10 conc.
IL-2
TNF-a
IL-12p70
IL-6
IL-12p40
IP-10
IL-13
Neg. Ctrl
Capture antibody

Does high expression of estrogen receptor and DNA damage signaling (γ-H2AX) indicate poor prognosis in female patients with resected Non-small Cell Lung Cancer?

Tatyana Zhukov*, Sangy Pottackal, Aleksandra Zajac, Rachna Kapoor, Dan Sheng Song, Wei Qian, Ambuj Kumar, Melvyn Tockman

Division of Cancer Prevention and Control – H. Lee Moffitt Cancer Center and Research Institute at the University of South Florida, 12902 Magnolia Drive, Tampa, FL 33612, USA

*Corresponding author: Dr TA Zhukov, PhD, PMIAC Department of Interdisciplinary Oncology, H. Lee Moffitt Cancer Center and Research Institute, University of South Florida, 12902 Magnolia Drive, Tampa, FL 33612, USA.

Tel: 813 745 1718; fax: 813 745 6525

E-mail: zhukovta@moffitt.usf.edu

Short running title: ER status and DNA damage signaling for gender-based prognosis in lung cancer

Summary

The lung cancer death rate in women has doubled over the past 25 years while the male lung cancer death rate has continued to decline. Although several lines of evidence suggest that women may be more susceptible to tobacco-induced lung cancer than men, definitive results related to disparity in lung cancer rates for men and women are lacking. We hypothesized that estrogen through interaction with estrogen receptors (ER) may modulate the DNA damage/repair signaling network and might be the reason for gender differences in the risk for lung cancer. Taking together, an actively functioning ER pathway and impaired DNA repair mechanism may underline a more malignant phenotypic behavior of female lung tumors. We therefore evaluated ER expression and γ-H2AX level (phosphorylated histone H2AX), a marker of DNA damage response, in resected tissue specimens from male and female lung cancers, and its relation to patient's survival. We detected ER and γ-H2AX with immunohistolabeling and assessed marker expression quantitatively at sub-cellular localization by applying novel computer-aided pathology diagnosis (CAPD) imaging modalities. We observed significant increase in expression of ERα, ERβ (nuclear) and γ-H2AX in malignant tumors compared to normal adjacent lung. To adjust for multiple statistical comparisons of discriminatory factors representing cell morphology and marker expression values, we used Bonferroni-Holm correction. The lung tumor specimens from women showed increased expression of ER and γ-H2AX, more prominently in NSCLC early stages. However, there was no gender difference in survival. Similar survival with earlier stage distribution may indicate a worse prognosis for women.

Keywords: estrogen receptor status; DNA damage response; lung cancer risk biomarkers; molecular imaging

Introduction

In 2006 lung cancer will account for 30 percent of all cancer death in the U.S [1]. In fact, nearly 80,000 new lung cancer cases in U.S. women were diagnosed in 2005 and this cancer was projected to kill approximately 68,000 U.S. women – more than breast and ovarian cancer combined. Previous studies have shown that about 90 percent of the sex difference in lung cancer incidence was attributable to gender differences in smoking habits [2]. Additionally, several studies have reported that women receive diagnoses at a younger median age, suggesting an increased susceptibility to the development of lung cancer [3]. A study by Henschke et al provided evidence related to the association of a given level of smoking and occurrence of lung cancer more often in women than men using baseline CT screening for lung cancer [4]. However, other studies have contradicted the findings of Henschke et al, and until now the effect of gender on the lung cancer risk associated with tobacco use remains unclear [5]. Although not definintive, some studies have shown that women who smoke appear to be at higher risk of developing small cell lung cancer than squamous cell lung cancer, whereas men who smoke have a similar risk for the two lung cancer histologic cell types [6,7]. Furthermore, women smokers are more likely to develop adenocarcinoma of the lung at a younger age, and estrogens may play a causative role in this phenomenon [8,9]. There is an evidence of genetic predisposition to lung cancer, especially in women aged ≤50 at the time of diagnosis [9]. More over, gender variation has also been studied in genes encoding carcinogen- metabolizing enzymes, including CYP1A1 that is involved in the metabolism of benzopyrene, a suspected tobacco-related lung carcinogen [10]. It is therefore conceivable that genetically determined variations in its activity modify individual susceptibility to lung cancer [11,12]. Among lung cancer patients, female smokers have been found to have higher levels of PAH-related DNA adducts and CYP1A1 gene expression in their normal lung tissue compared to male smokers [13]. A possible role of steroid hormones in these sex differences via interactions between aryl hydrocarbon receptor and estrogen receptor mediated cellular effects has been suggested, however reported data do not support the hypothesis of the role of estrogen in regulating the metabolic activation of polycyclic aromatic hydrocarbons in lung [14].

Despite the contradictory hypotheses related to gender bias in lung cancer rates, a potential unifying hypothesis exists for the explanation of sex differences in lung cancer presentation and susceptibility at the genetic and biochemical levels, and that relates to the estrogen and ER status. There is evidence supporting the hypothesis that early menopause decreases risk of adenocarcinoma of the lung, whereas hormone replacement therapy is associated with higher risk of lung cancer [15]. Recent studies suggest that estrogen receptor signaling pathways might play an important role in normal lung biology and in controlling the growth of lung cancer [15]. Estrogen may promote lung cancer through direct actions on pre-cancer or cancer cells, or through indirect action on lung fibroblasts [16]. Estrogen levels are often elevated in female lung cancer patients compared with women without lung cancer [17]; synthesis of estrogen might be increased focally in lung cancer cells [16]. Clinical analysis on ERα and ERβ expression in lung cancer so far has been contradictory [18,19].

We hypothesized that estrogen through interaction with estrogen receptors may modulate the DNA damage/repair signaling network and contribute to gender bias in lung cancer risk. Estrogen accelerates DNA synthesis, which may result in accumulation of DNA damage in the presence of an effective G2/M checkpoint. Loss of DNA repair or checkpoint function might lead to gender differences in lung tumor progression and contribute to aggressive malignant phenotypic behavior of female lung cancers. Consequently, the primary objectives of this study was to determine whether expression of estrogen receptors and a marker of DNA damage/repair signaling (γ-H2AX) are correlated to the incidence of lung cancer, and if there are any gender related differences in the expression of these markers and patient’s survival.

Materials and methods

Tissue specimens:

From the existing bank of surgically resected tumors, collected by the Moffitt Tissue Bank, we selected a set of lung cancer specimens from patients with no prior chemotherapy. From 19 lung cancer cases, we obtained snap-frozen samples of the tumors and corresponding sections of formalin fixed, paraffin embedded tumors, cut at 3 µm. We also obtained clinical information including gender, age, smoking history, tumor stage, and histologic type related to the specimens. Private identifiable patient information was removed from these records in accordance with IRB and HIPAA regulations. Slides of paraffin sections stained with H&E, that correspond with primary tumor were reviewed prior to inclusion according to established morphological criteria [20], and to assure presence of adjacent, uninvolved lung and pre-malignant lung lesions (peripheral AAH), as previously described [21].

Immunohistochemistry:

Detection of γ-H2AX in lung cancer tissue: we optimized an immunohistochemistry (IHC) staining using antigen retrieval and avidin/biotin blocking (DAKO) on formalin-fixed, paraffin-embedded freshly sectioned tissue. With antibody to γ-H2AX (1:700; Rabbit, polyclonal IgG, Upstate Biotechnology, Lake Placid, NY) applied overnight at 4ºC in a humid chamber, IHC was completed on a DAKO autostainer using Vector Elite – PX Rabbit detection and DAB. In the negative control (on lung tissue) primary antibody was omitted. When it was necessary, corresponding snap-frozen samples of the tumors were used to confirm marker expression.

Estrogen receptor immunohistologic staining: consecutive sections from the same tissue blocks described above were used for staining with antibodies to ERα. and ERβ. We optimized procedure for immunostaining of ER-β with mouse monoclonal antibody to Oestrogen R-beta 1 (Serotec, UK). Deparaffinized sections underwent microwave (700 W) antigen retrieval in citrate buffer, blocking with avidin/biotin and then incubation with primary antibody (1:400) applied for 24 hours. IHC was completed on a DAKO autostainer using Vector Elite – PX Mouse detection. ERα staining on tissue sections was performed after antigen retrieval, and application of the standard Ventana test (with clone 6F11) routinely used in the diagnostic histopathology laboratories of the Moffitt Cancer Center and the Tampa General Hospital. The positive control for ERα and ERβ were breast cancer and uterus tissues; in the negative control primary antibodies were replaced with PBS, following secondary antibody detection.

Imaging

Bright field microscopy: A Nikon E600 light microscope equipped with CCD image capture, using Spot Advanced software (Diagnostic Instruments Sterling Heights, MI) were utilized in this study. Examples of processed images are illustrated in Figures 1, 2, which show regions of interest (ROI) on tissue sections (usually defined as tumor front in malignant lung, and as random region in normal lung alveolar-bronchial/bronchiolar septa), immunostained and imaged for quantitative assessment of descriptive cell features.

Computer-Aided Pathology Diagnosis (CAPD) image analysis:

Images of tissue sections were processed with CAPD – software modules that analyze digital 12-bit color images of stained sections for obtaining mathematical means of cell morphology features, and to interpret target marker expression intensity parameters that include color separation on spectral components. The basic modules of CAPD system are structured as follows: (a) nonlinear, multistage and adaptive filtering for image noise suppression and for artifact reduction as required for implementation of high order wavelet transforms that are sensitive to noise [22], (b) Fragmentary Window Filter (FWF) designed for using only a fraction of pixels in the filter window to define a specific geometric pattern and emphasizing the geometric property of the pattern over its contrast [23], (c) multiresolution and multi-orientation wavelet for improved feature extraction using the unique properties of wavelet transforms, in standard and tree-structured forms, implemented on filter banks to preserve image details [24], (d) Segmentation algorithm by using Multivariate Fuzzy C-Means clustering (MFCM), (e) genetic algorithm for feature ranking/selection [25], and (f) Single and multistage Neural Networks (NN's) with significantly increased convergence speed for more efficient classification [26]. In this study a novel, fully automatic, and highly efficient method was developed for the optimization of CAPD system, including algorithm module optimization, histological and cytological knowledge structure optimization [27]. A statistic test for multiple comparisons was used to validate the CAPD discriminatory criteria.

Western blotting: To confirm IHC findings on tested markers in resected lung specimens we conducted western blot analysis. Proteins were extracted from microdissected tumor/normal tissue and cell samples in 1 ml of Lysis buffer [9.5 M Urea, 2% CHAPS, 0.5% DTT] and incubated on ice for 20 minutes. During the incubation stage, the lysate was sonicated using a probe sonicator twice for 5 seconds with 1 minute on ice in between bursts. Following incubation, the lysate was centrifuged at 10000 x g for 20 minutes at 4° C, and the supernatant was transferred to another tube. Protein concentration was measured using the Bio-rad Protein Assay kit. For detection of γ-H2AX, equal amounts of cell lysate (30 ug/ml protein) were separated by size on 12% SDS-PAGE and transferred to Immobilon-P (PVDF) membranes, following detection steps. Anti-phospho-H2A.X (ser139) (Upstate cell signaling solutions, Lake Placid, NY) at 1:5000 dilution was used as the detector antibody.

For identification of ER, equal amounts of cell lysate were separated on a 8% SDS-PAGE and transferred to PVDF membranes. ERα (1:800; clone D-12) and ERβ (1:500) detector antibody (Santa Cruz Biotechnology, CA), were applied for 72 hours.

Cell lines and culture conditions: NSCLC cell lines, H23, Calu 3 (adenocarcinomas); A549 (bronchioloalveolar carcinoma, BAC), Sk-Mes 1, Sk-Lu 1, H157 (squamous cell carcinomas) were purchased from American Type Culture Collection (Rockville, MD). All cell lines were maintained in growth medium RPMI 1640 supplemented with 10% fetal bovine serum, 2 mM L-glutamine (Life Technologies), 10 mM HEPES and antibiotics. H69 SCLC cell culture was maintained in modified RPMI 1640 medium as described in [28]. Cultures were maintained in humidified incubators at 37º C in an atmosphere of 5% CO_2 in air. Non-neoplastic cultured cell lines, normal human pulmonary fibroblasts, NIH 373, were used as control to malignant cells and maintained in DMEM media, supplemented with 10% heat inactivated fetal bovine serum, 1 mM glutamine, and antibiotics. The breast carcinoma cell line MCF-7, positive for estrogen receptor, was used as positive control in our study. The MCF-7 cell line which expresses the wild-type *p53*, was grown in DMEM media, supplemented with 10% heat inactivated fetal bovine serum, 1 mM glutamine, and antibiotics. For each experimental estrogen–free condition, cells were cultured 48 h before experiments in phenol-red free RPMI medium 1640 (Sigma) with 1% dextran-coated, charcoal-treated FBS.

Statistical analysis: Using SAS (version 9.1), descriptive statistics were generated for demographic and clinical variables. Paired T-test was applied to find a set of cell descriptive features (morphology parameters and markers expression values) that were independently predictive of tumor/normal status. To adjust for multiple statistical comparisons of CAPD discriminatory factors representing cell morphology and target marker expression values, we used Bonferroni-Holm correction for tumor/normal discrimination. Kaplan-Meier survival curve was plotted to assess significant difference in survival for men and women with lung cancer. With respect to levels of the potential markers discussed above those found to be significant ($P < 0.05$) were used to determine whether or not profiles of any subset of the markers can distinguish cancer from the non-cancer cells in clinical samples.

Results

Our tissue sample consisted of nineteen patients (10 female and 9 male, age range 56 to 83). We stratified patients' age into two groups (≤70 and > 70 years) based on the median. Characteristics of patients have been shown in detail in table 2.

Thirty percent of females had a mild smoking history (<1 PPD), 20% had a moderate (>1-2 PPD) smoking history and 50% never smoked. In male patients 44% had a mild and 56% had moderate. None of the male patients had a no-smoking history. Sixteen cases represented Non-Small Cell Lung Cancer (NSCLC), including 4 Adenocarcinoma, 4 Squamous cell carcinoma, 4 Bronchioloalveolar carcinoma and 4 Large cell undifferentiated carcinoma of the lung. In addition we included three cases of SCLC (Small Cell Lung Cancer). Follow-up information was available for all nineteen patients, two NSCLC male patients are still alive with no evidence of lung cancer at 97 months and 60 months, respectively. However, all ten female patients died of disease.

Kaplan-Meier survival curves were plotted to assess the prognosis of lung cancer patients. Overall survival and median survival according to gender in nineteen lung cancer cases do not show statistically significant difference (Figure 3 and table 3). Although lung cancer cases among the women were predominantly diagnosed at an earlier stage (I, IB) compared to men (3A, 4), (Table 2), still survival was not different in males and females. An early stage detection with no survival advantage may indicate a worse prognosis for women with lung cancer at their late menopause and post menopause age.

Also, it is important to note that among female lung cancer patients in this group 50% were reported as never smokers. In contrast, all male patients were former smokers. As stated earlier, overall survival was not different in both groups (male and females) despite the fact that males had higher rates of smoking history, and were diagnosed at later stages. However, in all NSCLC patients expression level of targeted markers (ERs and γ-H2AX) was elevated (Table 1). Nevertheless, it was higher in female NSCLC patients.

Immunohistochemical analysis of tumor tissue specimens

The immunohistochemistry for ER and γ-H2AX was performed on a set of nineteen lung tumor and normal adjacent tissue specimens. With microscopy we observed increased expression of γ-H2AX with predominately nuclear staining in squamous cell carcinomas, and nuclear and cytoplasmic staining in cases of adenocarcinoma of the lung. Of great interest is the strong nuclear antigen expression in cells from an AAH lesion (potentially pre-malignant lesion of the lung) with baseline detectable γ-H2AX staining in morphologically normal adjacent tissue (Figure 1). The same NSCLC cases showed elevated expression of ERβ, and focal expression of ERα. Low expression of γ-H2AX was noted in cases of SCLC.

The IHC results on marker expression assessment in frankly malignant and normal lung cells with our CAPD image analysis system [27] are summarized in Table 1. As we can see, extracted features represent cell morphology and marker expression at sub-cellular locations. Paired T- test was applied to CAPD mean values to find a set of features that were predictive of lung cancer. To adjust for multiple statistical comparisons, we used Bonferroni-Holm correction. With level of significance ($p<0.05$) we found, that expression values of γ-H2AX were increased in nucleus of lung tumor cells compared to normal cells (the best discriminatory variable for this marker was found – spectral Red, $p<0.0001$). Analysis of expression intensity of ERα and

ERβ showed significant increase for both types of receptor in nucleus of malignant lung cells (spectral Red, Green, Luminance, $p<0.0001$; and spectral Red, $p=0.003$, respectively). Selected CAPD descriptive features of cell morphology showed remarkable changes in nucleus of malignant lung cells compared to normal cells (area, convex area; roundness, skeleton length, $p<0.005$).

Analysis of gender bias in targeted markers expression revealed (without correction for multiple comparisons due to exploratory nature of the gender-based discrimination in this study) that CAPD numerical values of cell descriptive features in normal lung did not show significant differences for γ-H2AX and ERα; however values of ERβ expression were slightly increased in female normal lung cells (nuclear, spectral Red, $p=0.0184$). CAPD features of cell morphology did not show any gender differences in normal lung cells. When we analyzed the same parameters in lung tumor tissue we noticed changes in γ-H2AX level and ERα increase of expression in female lung cancer cells. Marker intensity expression analysis revealed that γ-H2AX is increased in female tumor cells, as shown by value of nuclear spectral Blue ($p=0.0386$); whereas for ERα, CAPD computed spectral intensity values were increased at extranuclear location (cytoplasm Luminance, $p=0.0106$) and in nucleus (Saturation, $p=0.0140$). Interesting to note that expression level of targeted markers (ERα and γ-H2AX) was not correlated with smoking status of female patients; although, association between gender and smoking history was not significant in this pilot group. These data support our hypothesis that gender-based differences in lung cancer risk and prognosis are related to the status of estrogen receptor activity and level of DNA damage/ repair response.

Western blot analysis

Immunoblotting confirmed IHC/CAPD results on marker level in malignant lung compared to normal lung tissue. Figure 4 illustrates that lung cancer cells show elevated level of γ-H2AX compared to microdissected cells from the normal adjacent lung tissue. We noticed that in lane 3 (case 2137) expression of γ-H2AX is the most intensive among shown on this gel (Figure 4B). This case represents a white non-Hispanic, 74 years old female lung cancer patient, diagnosed with an early stage second primary adenocarcinoma of the lung. In contrast, lanes 1 and 5 show significantly lower expression of γ-H2AX representing male lung tumors, clearly indicating gender differences in the level of this marker. In concordance, we found increased expression of ERα on immunoblot from female lung cancer tissue compared to male lung tumor, and above the level of expression in normal tissue (Figure 5).

Detection of these markers in cultured NSCLC cell lines (A549, H23, Sk-Mes 1, Calu 3) showed elevated level of γ-H2AX in malignant cells compared to non-malignant cells (pulmonary fibroblasts), Figure 4. Interesting to note that relative level of γ-H2AX detected with IHC in SCLC cases was low (observed only in one out of three cases), and was not detected by immunoblotting in SCLC cell line H69 (Figure 4C). This observation might be explained by DNA damage signaling pathway through H2AX is not actively expressed in highly aggressive lung cancer types. However, on the other hand, increased γ-H2AX level is observed in early stage NSCLC, and even in second primary lung cancer, as we have noted. Together with increased estrogen receptor expression, these findings may reflect conditions of activated DNA damage/repair signaling network early in transformation. Therefore, γ-H2AX and ERs expression may be viewed potentially as informative markers of cancer risk and prognosis in patients with resected NSCLC.

Discussion

We hypothesized that estrogen through binding with estrogen receptors and ER signaling pathways activation may modulate the DNA damage/repair response, and be responsible for the gender bias in lung cancer risk and progression. Estrogen accelerates DNA synthesis [29], which may result in accumulated DNA damage in the presence of an effective G2/M checkpoint. Impaired DNA repair or checkpoint function might lead to gender differences in lung cancer risk [30]. The purpose of the present paper was to determine whether expression of estrogen receptors and marker of DNA damage/repair signaling pathway (γ-H2AX) are correlated to lung cancer and associated with gender bias. Additionally we believe that precise quantification of level of ER expression and γ-H2AX will allow subsequent validation assessment of their feasibility as potential markers for lung cancer growth and progression in female as well as male lung cancer patients. We detected markers with immunohistolabeling on resected clinical lung tumor tissue specimens, and assessed marker expression quantitatively at sub-cellular localizations by applying novel CAPD imaging modalities [27].

Limited data are available about estrogen receptor pathways in lung cancer [16]. Two estrogen receptors, ERα and ERβ have been identified, which are encoded by separate genes, and function as ligand-activated transcription factors [31]. Whether ERα or ERβ mediate the ligand-dependent transcriptional response of NSCLC cells to estrogen, remains controversial. ERβ immunostaining was found more often in adenocarcinomas, compared to squamous cell carcinomas of the lung [32]. In lung cancer cells, estrogen can directly stimulate the transcription of estrogen-responsive genes along the EGF pathway [33]. The presence ERβ in the lung, and the fact that estrogen could stimulate lung tumor growth is intriguing, and opposite from ERβ inhibitory effect in the carcinogenesis of other tissues, as mammary and prostate [34]. The mechanisms of these observations are unknown and may occur through the function of ERα.

DNA damage/repair signaling: It has been shown that phosphorylation of histone H2AX is linearly correlated to the number of DSB [35]. Immunofluorescence studies have shown that several proteins involved in DNA repair including BRCA1, BRCA2, Rad51 and Mre11 are recruited to sites of γ-H2AX [36]. γ-H2AX seems to localize specifically at sites of damage and it has been recognized as one of the earliest markers of DNA damage [37]. This provides a basis for the use of γ-H2AX as biomarker for cancer [38]. Recent data suggest that H2AX is a tumor suppressor gene, raising the possibility that mutations or changes in levels of H2AX may be causally involved in cancer development [39]. It was proposed that ATR/ATM-regulated DNA damage checkpoint (phosphorylated kinases ATM and Chk2, γ-H2AX and p53) might become activated in the early stages of human tumorigenesis that delay or prevent cancer, if this checkpoint is not compromised [40,41]. Regarding the fate of γ-H2AX as indicator of DNA damage signaling, it was suggested that the dephosphorylation of γ-H2AX is necessary for the termination of the checkpoint signaling after the DNA repair process is complete, but is not necessary for the resolution of the repair process [42,43].

In the present study we observed significantly ($p<0.005$) increased expression of γ-H2AX, ERβ, and focal expression of ERα in lung tumors compared to normal adjacent tissue for all NSCLC patients. CAPD image analysis generated mathematical values of descriptive features extracted from imaged cells over regions of interest on lung tissue sections (with Bonferroni-Holm correction we adjusted CAPD values for multiple statistical comparisons). As discovery by nature (not definitive), we noted gender differences in marker expression - the lung cancer specimens from women showed elevated ER and γ-H2AX expression, and an earlier stage distribution compared to men although there was no gender difference in survival. Similar survival of earlier stages cases may indicate a worse prognosis for women. Currently accepted clinical dogma states that survival from lung cancer is related to stage at presentation. However, in our pilot study there was no difference in survival rates in female and male lung cancer patients according to stage of the disease. The reasons for this difference could not be verified in the current study due to small number of cases. Nevertheless, the limited data still showed the correlation between over-expression of estrogen receptors and DNA damage [44], and the risk of NSCLC in female lung cancer patients.

Harris et al reported (a) that gender constitutes a major prognostic factor in SCLC and is especially useful as a predictor for long-term survival, and (b) that the favorable prognostic value of the female sex is restricted to younger patients [45]. Our results obtained on 19 lung cancer patients in the present study are in concordance with Harris. We observed poor prognoses for women aged ≥60 years diagnosed with early stages of the disease, experiencing survivals comparable to men diagnosed with advanced stages (median survival time 32 mo vs. 28 mo for female and male, respectively). In regards to the role of smoking and lung cancer risk in females, it is important to note here studies of Henschke et al [4] that found association of a given level of smoking and occurrence of lung cancer more often in females than in males using baseline CT screening for lung cancer. However, other studies [5] have contradicted the findings of Henschke et al, and the questions on the effect of gender on the lung cancer risk associated with tobacco use still under investigation [46] and acquire new directions [47].

The observation in our study that half of the female lung cancer patients were non smokers compared to male patients who were all smokers lead us to the existence of additional factors in the poor prognosis of female NSCLC patients, which are not exactly attributable to smoking status and early tumor stage. We recognize that the association of overexpressed ERα, ERβ with increased level of γ-H2AX in female lung cancer with worse prognosis must now be validated in an independent set of lung tumor specimens. Analytic validation of these markers must then be followed by clinical validation to demonstrate that markers of ER activation and DNA damage/repair response may serve as prognostic biomarkers identifying high-risk patients, specifically in women diagnosed with early stage lung cancer.

Acknowledgements

We thank the Biostatistics Core at the H. Lee Moffitt Cancer Center and Research Institute for their invaluable input in statistical analysis of our data.

This work was supported in part by National Cancer Institute grants No.6302-295-LO and No. SR33CA099046-01A1.

Conflict of interest statement

None declared.

Figure 1.

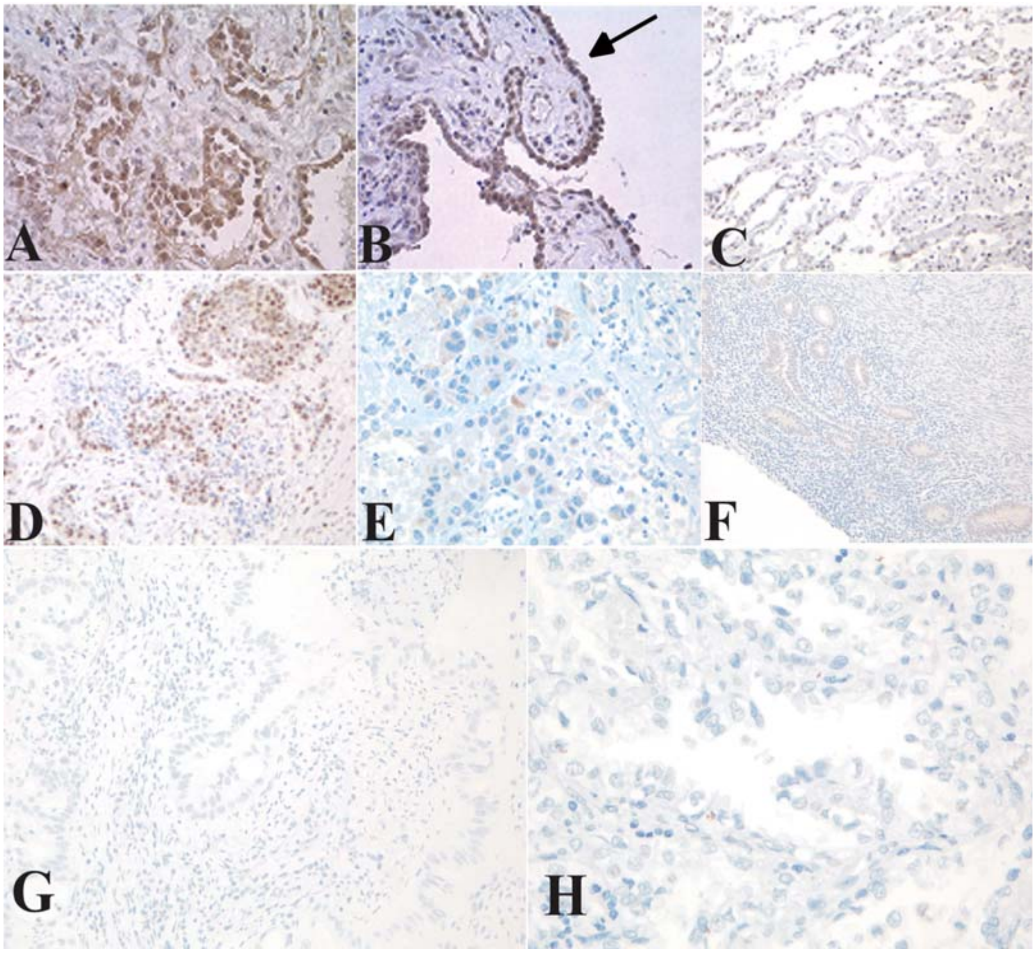


Figure 2.

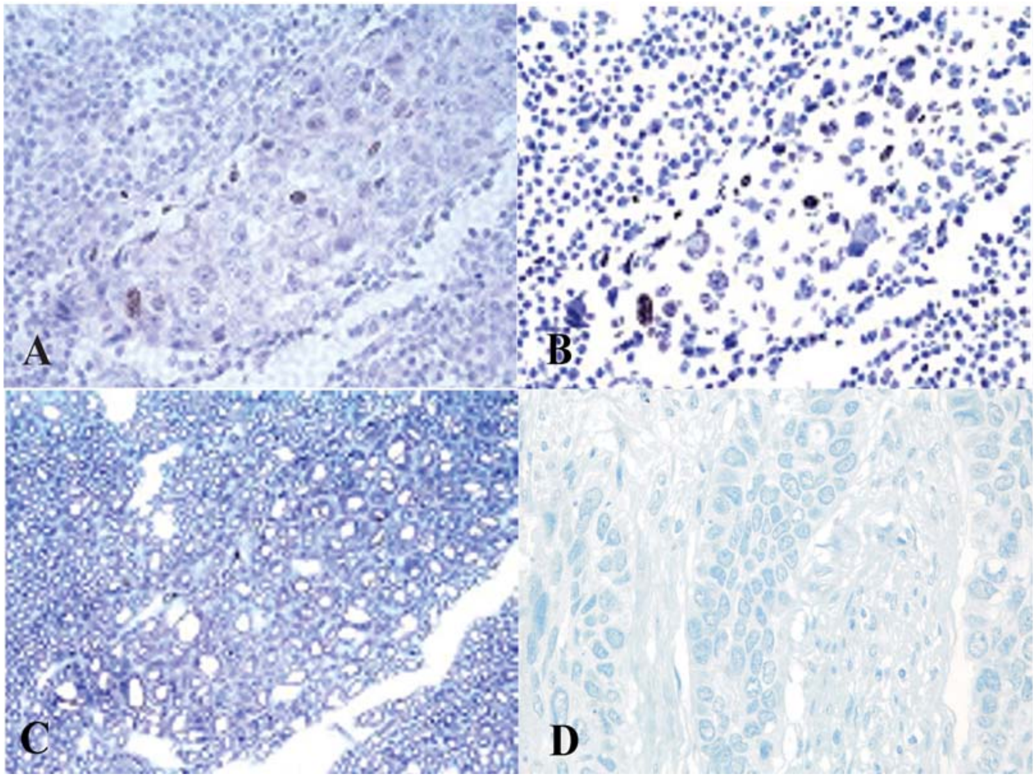


Figure 3

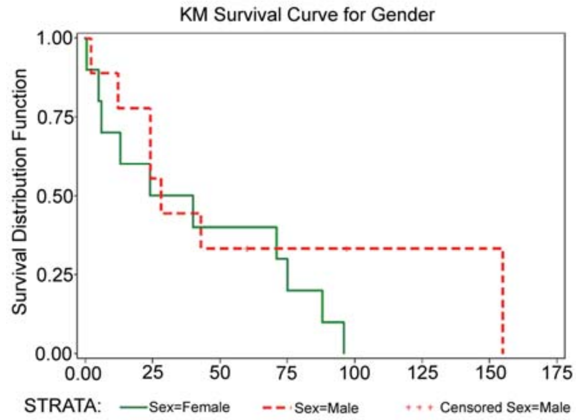
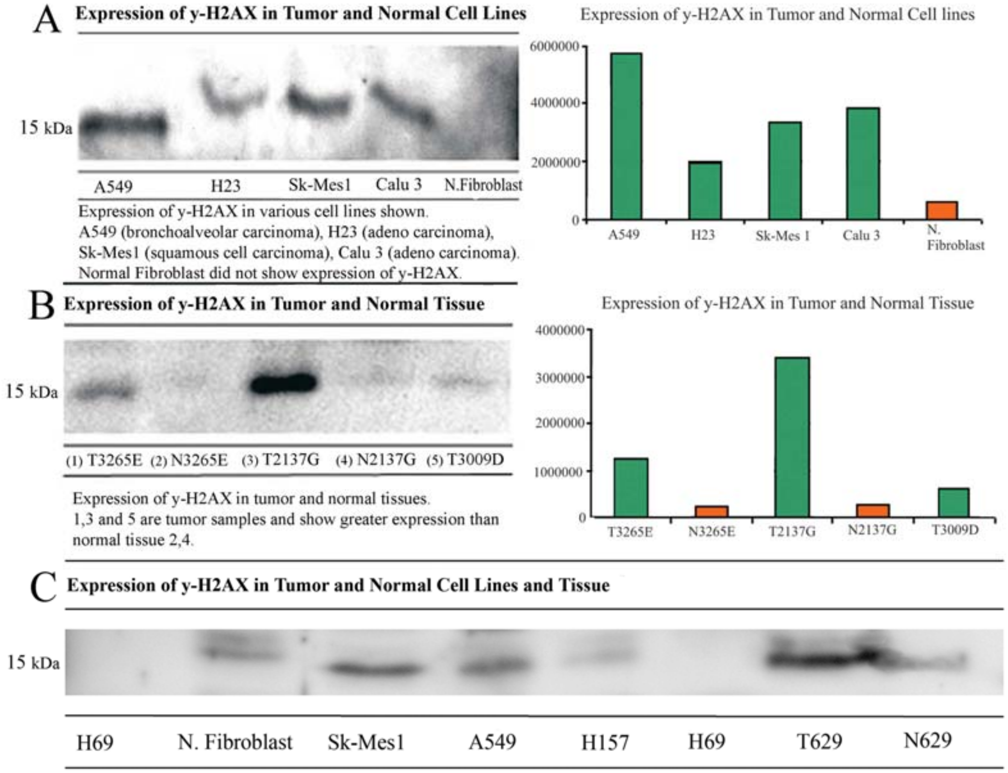

KM Survival Curve for Gender
Survival Distribution Function
1.00
0.75
0.50
0.25
0.00
0.00
25
50
75
100
125
150
175
STRATA:
Sex=Female
Sex=Male
Censored Sex=Male

Figure 4



Expression of y-H2AX in various cell lines shown.
A549 (bronchoalveolar carcinoma), H23 (adeno carcinoma),
Sk-Mes1 (squamous cell carcinoma), Calu 3 (adeno carcinoma).
Normal Fibroblast did not show expression of y-H2AX.

Expression of y-H2AX in tumor and normal tissues.
1,3 and 5 are tumor samples and show greater expression than
normal tissue 2,4.

Note: In A and B - Anti-phospho-H2AX (ser 139) (Upstate cell signaling solutions, Lake Placid, NY) was used at 1:5000 dilution as the primary antibody for membrane; in C - Anti-phospho-H2AX was used at 1:1000 dilution.

Figure 5

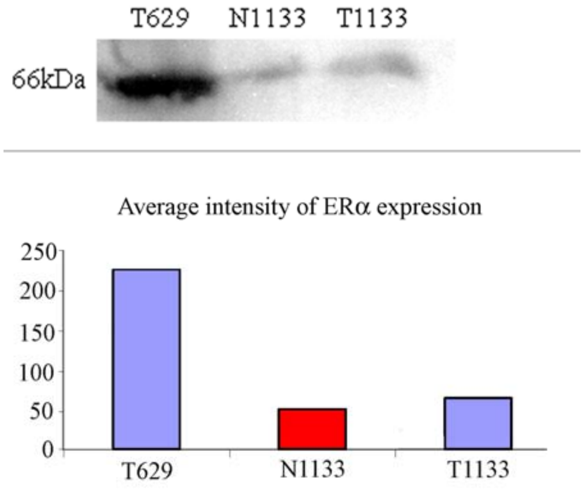

T629
N1133
T1133
66kDa
Average intensity of ERα expression
250
200
150
100
50
0
T629
N1133
T1133

Table 1. Tumor versus normal tissue expression of γ-H2AX and Estrogen Receptor- α and β in nineteen lung tumors and paired adjacent “normal’ tissue

CAPD Features		**γ – H2AX expression**		**ER-α expression**		**ER-α expression**		**ER-β expression**		**ER-β expression**	
		localization: Nuclear		**localization: Cytoplasm**		**localization: Nuclear**		**localization: Cytoplasm**		**localization: Nuclear**	
Normal/ Tumor		Mean (STD) of differences	Paired T-test P- value	Mean (STD) of differences	Paired T-test P- value	Mean (STD) of differences	Paired T-test P- value	Mean (STD) of differences	Paired T-test P- value	Mean (STD) of differences	Paired T-test P- value
Morphological Features	Area	-0.350 (0.28)	*0.0049	0.002 (0.34)	1.0000	-0.435 (0.14)	*<.0001	0.135 (0.26)	1.0000	-0.335 (0.20)	*0.0002
	Convex Area	-0.341 (0.29)	*0.0096	-0.070 (0.29)	1.0000	-0.428 (0.14)	*<.0001	0.084 (0.26)	1.0000	-0.330 (0.20)	*0.0003
	Formfactor	-0.031 (0.04)	0.3712	0.116 (0.10)	*0.0138	-0.019 (0.05)	1.0000	0.080 (0.08)	*0.0352	-0.011 (0.03)	1.0000
	Roundness	-0.079 (0.07)	*0.0096	0.017 (0.07)	1.0000	-0.077 (0.08)	*0.0430	0.019 (0.04)	1.0000	-0.038 (0.05)	0.1406
	Skeleton Length	-0.176 (0.20)	0.0741	-0.055 (0.23)	1.0000	-0.251 (0.10)	*<.0001	0.047 (0.17)	1.0000	-0.165 (0.13)	*0.0036
Color and Intensity Features	Red	-0.297 (0.15)	*<.0001	-0.060 (0.07)	0.0741	-0.393 (0.21)	*<.0001	0.001 (0.05)	1.0000	-0.205 (0.16)	*0.0030
	Green	-0.198 (0.18)	*0.0180	-0.012 (0.06)	1.0000	-0.166 (0.09)	*<.0001	0.029 (0.06)	1.0000	-0.103 (0.29)	1.0000
	Blue	0.013 (0.18)	1.0000	0.013 (0.08)	1.0000	-0.049 (0.05)	0.0600	0.065 (0.08)	0.1575	0.068 (0.20)	1.0000
	Hue	0.021 (0.07)	1.0000	0.032 (0.27)	1.0000	0.016 (0.02)	0.2074	0.412 (0.51)	0.1476	-0.049 (0.06)	0.2145
	Saturation	0.184 (0.20)	0.0574	0.058 (0.24)	1.0000	0.053 (0.13)	1.0000	-0.116 (0.21)	0.9884	0.186 (0.25)	0.2542
	Luminance	-0.149 (0.15)	*0.0430	-0.018 (0.05)	1.0000	-0.141 (0.08)	*<.0001	0.031 (0.06)	1.0000	-0.067 (0.19)	1.0000
	Intensity Standard Deviation	0.048 (0.09)	1.0000	0.028 (0.09)	1.0000	-0.094 (0.13)	0.2464	0.003 (0.14)	1.0000	0.066 (0.09)	0.3270

Notice: Log-transformed Bonferroni Holm adjusted P-values; * Level of significance = 0.05

PAGE LEFT BLANK

Table 2 Demographic and clinical characteristics for the study subjects by gender: (N=19)

	Male (n=9)	**Female (n=10)**
	N (%)	**N (%)**
Age in years		
≤ 70	3 (33.3)	6 (60.0)
> 70	6 (66.7)	4 (40.0)
Smoking History		
Mild	4 (44.4)	3 (30.0)
Moderate	5 (55.6)	2 (20.0)
None-Never Smoked	0	5 (50.0)
AJCC Stage		
1	0	3 (30.0)
1B	0	2 (20.0)
2	1 (11.1)	1 (10.0)
2B	1 (11.1)	1 (10.0)
3A	4 (44.4)	2 (20.0)
4	3 (33.3)	1 (10.0)
Vital Status		
Alive	2 (22.2)	0
Dead	7 (77.8)	10 (100.0)

Table 3 Survival characteristics

Gender	**Frequency (%)**	**Median Survival Time (mths)**	**Log-Rank p-value=0.3203**
F	10 (53)	32	
M	9 (47)	28	

References

[1] Jemal A, Murray T, Samuels A, Ghafoor A, Ward E, Thun MJ. Cancer statistics, 2003. CA Cancer J Clin, 2003. 53(1): p. 5-26.

[2] Tong L, Spitz MR, Fueger JJ, Amos CI. Lung carcinoma in former smokers. Cancer, 1996. 78(5): p. 1004-10.

[3] Payne S. 'Smoke like a man, die like a man'?: a review of the relationship between gender, sex and lung cancer. Soc Sci Med, 2001. 53(8): p. 1067-80.

[4] Henschke CI and Miettinen OS. Women's susceptibility to tobacco carcinogens. Lung Cancer, 2004. 43(1): p. 1-5.

[5] Fu JB, Kau Y, Severson RK, Kalemkerian GP. Lung cancer in women: analysis of the national Surveillance, Epidemiology, and End Results database. Chest, 2005. 127(3): p. 768-77.

[6] Kazerouni N, Alverson CJ, Redd SC, Mott JA, Mannino DM. Sex differences in COPD and lung cancer mortality trends--United States, 1968-1999. J Womens Health (Larchmt), 2004. 13(1): p. 17-23.

[7] de Perrot M, Licker M, Bouchardy C, Usel M, Robert J, Spiliopoulos A. Sex differences in presentation, management, and prognosis of patients with non-small cell lung carcinoma. J Thorac Cardiovasc Surg, 2000. 119(1): p. 21-6.

[8] Baldini EH and GM Strauss. Women and lung cancer: waiting to exhale. Chest, 1997. 112(4 Suppl): p. 229S-234S.

[9] Sellers TA, Potter JD, Bailey-Wilson JE, Rich SS, Rothschild H, Elston RC. Lung cancer detection and prevention: evidence for an interaction between smoking and genetic predisposition. Cancer Res, 1992. 52(9 Suppl): p. 2694s-2697s.

[10] Dresler CM, Fratelli C, Babb J, Everley L, Evans AA, Clapper ML. Gender differences in genetic susceptibility for lung cancer. Lung Cancer, 2000. 30(3): p. 153-60.

[11] Berge G, Mollerup S, OVrebo S, Hewer A, Phillips DH, Eilertsen E, Haugen A. Role of estrogen receptor in regulation of polycyclic aromatic hydrocarbon metabolic activation in lung. Lung Cancer, 2004. 45(3): p. 289-97.

[12] Vineis P, Veglia F, Benhamou S, Butkiewicz D, Cascorbi I, Clapper ML, Dolzan V, Haugen A, Hirvonen A, Ingelman-Sundberg M, Kihara M, Kiyohara C, Kremers P, Le Marchand L, Ohshima S, Pastorelli R, Rannug A, Romkes M, Schoket B, Shields P, Strange RC, Stucker I, Sugimura H, Garte S, Gaspari L, Taioli E. CYP1A1 T3801 C polymorphism and lung cancer: a pooled analysis of 2451 cases and 3358 controls. Int J Cancer, 2003. 104(5): p. 650-7.

[13] Mollerup S, Ryberg D, Hewer A, Phillips DH, Haugen A. Sex differences in lung CYP1A1 expression and DNA adduct levels among lung cancer patients. Cancer Res, 1999. 59(14): p. 3317-20.

[14] Shriver SP, Bourdeau HA, Gubish CT, Tirpak DL, Davis AL, Luketich JD, Siegfried JM. Sex-specific expression of gastrin-releasing peptide receptor: relationship to smoking history and risk of lung cancer. J Natl Cancer Inst, 2000. 92(1): p. 24-33.

[15] Taioli E and Wynder EL. Re: Endocrine factors and adenocarcinoma of the lung in women. J Natl Cancer Inst, 1994. 86(11): p. 869-70.

[16] Stabile LP and Siegfried JM. Estrogen receptor pathways in lung cancer. Curr Oncol Rep, 2004. 6(4): p. 259-67.

[17] Tiutiunova AM, Chirvina ED, Mironenko TV, Kartashov SZ, Luntovskaia VA. Hormonal balance in women with lung cancer and its changes after combined treatment. Vopr Onkol, 1986. 32(4): p. 26-30.

[18] Kawai H, Ishii A, Washiya K, Konno T, Kon H, Yamaya C, Ono I, Minamiya Y and Ogawa J. Estrogen receptor alpha and beta are prognostic factors in non-small cell lung cancer. Clin Cancer Res, 2005. 11(14): p. 5084-9.

[19] Pietras J.R, Marquez D, Chen Hsiao-Wang, Tsai E, Weinberg O, Fishbein M. Estrogen and growth factor receptor interactions in human breast and non-small cell lung cancer cells. Steroids, 2005. 70: p. 372-381

[20] Brambilla E, Travis WD, Colby TV, Corrin B, Shimosato Y. The new World Health Organization classification of lung tumours. Eur Respir J, 2001. 18(6): p. 1059-68.

[21] Zhukov TA, Johanson RA, Cantor AB, Clark RA, Tockman MS. Discovery of distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. Lung Cancer, 2003. 40(3): p. 267-79.

[22] Qian, W., Clarke L.P., Kallergi M., and Clark R.A. Tree structured nonlinear filters in Digital mammography. IEEE Trans. Med. Imag. March 1994, pp25-37.

[23] Mao F, Qian W, Gaviria J, Clarke LP. Fragmentary window filtering for multiscale lung nodule detection. Acad Radiol. 1998;5:306-311.

[24] Qian W, Kallergi M. Clarke LP, Li, HD, Clark RA, Silbiger, ML. Tree-structured nonlinear filter and wavelet transform for microcalcification segmentation in digital mammography. Med. Phys. 1995;22(8):1247-1254.

[25] Shen Y, Sankar R, Qian W, Sun XJ, Song DS. Fuzzy Image Segmentation For Lung Nodule Detection. Proceedings of IEEE, 2003 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 03), 6-10 April 2003; Hong Kong.

[26] Qain W, L.H. Li, and L.P. Clarke. Image feature extraction for mass detection using digital mammography: Influence of wavelet analysis. Medical Physics, 1999. 26(3): p. 402-408.

[27] Qian W, Zhukov TA, Song DS, Tockman MS. Computerized Analysis of Cellular Features and Biomarkers for Cytological Diagnosis of Early Lung Cancer. Accepted for publication in Journal of Analytical and Quantitative Cytology and Histology, 2006. In Press.

[28] Gazdar AF, Carney DN, Russell EK, Sims HL, Baylin SB, Bunn PA Jr, Guccion JG, Minna JD; Establishment of continuous, clonable cultures of small-cell carcinoma of lung which have amine precursor uptake and decarboxylation cell properties. Cancer Res, 1980. 40(10): p. 3502-7.

[29] Osborne CK, Schiff R. Estrogen-receptor biology: continuing progress and therapeutic implications. J Clin Oncol, 2005. 23(8): p. 1616-22.

[30] Wei Q, Cheng L, Amos CI, Wang LE, Guo Z, Hong WK, Spitz MR. Repair of tobacco carcinogen-induced DNA adducts and lung cancer risk: a molecular epidemiologic study. J Natl Cancer Inst, 2000. 92(21): p. 1764-72.

[31] Petterson K, Gustafsson JA. Role of estrogen receptor beta in estrogen action. Annu Rev Physiol, 2001. 63: p. 165-92.

[32] Omoto Y, Kobayashi Y, Nishida K, Tsuchiya E, Eguchi H, Nakagawa K, Ishikawa Y, Yamori T, Iwase H, Fujii Y, Warner M, Gustafsson JA, Hayashi SI. Expression, function, and clinical implications of the estrogen receptor beta in human lung cancers. Biochem Biophys Res Commun, 2001. 285(2):340-7.

[33] Stabile LP, Lyker JS, Gubish CT, Zhang W, Grandis JR, Siegfried JM. Combined targeting of the estrogen receptor and the epidermal growth factor receptor in non-small cell lung cancer shows enhanced antiproliferative effects. Cancer Res, 2005. 65(4): p. 1459-70.

[34] Roger P, Sahla ME, Makela S, Gustafsson JA, Baldet P, Rochefort H. Decreased expression of estrogen receptor beta protein in proliferative preinvasive mammary tumors. Cancer Res, 2001. 61(6): p. 2537-41.

[35] Hatch CL, W.M. Bonner, and E.N. Moudrianakis. Minor histone 2A variants and ubiquinated forms in the native H2A:H2B dimer. Science, 1983. 221(4609): p. 468-70.

[36] Ward IM and Chen J. Histone H2AX is phosphorylated in an ATR-dependent manner in response to replicational stress. J Biol Chem, 2001. 276(51): p. 47759-62.

[37] Rogakou EP, Boon C, Redon C, Bonner WM. Megabase chromatin domains involved in DNA double-strand breaks in vivo. J. Cell Biol, 1999. 146: 905-16.

[38] Monteiro AN, Zhang S, Phelan CM, Narod SA. Absence of constitutional H2AX gene mutations in 101 hereditary breast cancer families. J Med Genet, 2003. 40(4): p. e51.

[39] Burma S, Chen BP, Murphy M, Kurimasa A, Chen DJ. ATM phosphorylates histone H2AX in response to DNA double-strand breaks. J Biol Chem, 2001. 276(45): p. 42462-7.

[40] Bartkova J, Horejsi Z, Koed K, Kramer A, Tort F, Zieger K, Guldberg P, Sehested M, Nesland JM, Lukas C, Omtoft T, Lukas J, Bartek J. DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. Nature, 2005. 434(7035): p. 864-70.

[41] Bassing CH and Alt FW. The cellular response to general and programmed DNA double strand breaks. DNA Repair (Amst), 2004. 3(8-9): p. 781-96.

[42] Rios-Doria J, Fay A, Velkova A, Monteiro AN. DNA damage response: determining the fate of phosphorylated histone H2AX. Cancer Biol Ther, 2006 Feb;5(2):142-4.

[43] Keogh MC, Kim JA, Downey M, Fillingham J, Chowdhury D, Harrison JC, Onishi M, Datta N, Galicia S, Emili A, Lieberman J, Shen X, Buratowski S, Haber JE, Durocher D, Greenblatt JF, Krogan NJ. A phosphatase complex that dephosphorylates gammaH2AX regulates DNA damage checkpoint recovery. Nature. 2006; 439(7075): p. 497-501.

[44] Zhukov TA, Pottackal S, Monteiro A, Qian W, Song DS, Cantor AB, Sellers TA, Tockman MS. Novel Lung Cancer Biomarkers: γ-H2AX (Marker of DNA Damage Response) and Estrogen Receptors Pathways: Is there any crosstalk? In Proceedings Frontiers in Cancer Prevention Research, Fourth Annual AACR international Conference, Baltimore, MD. 2005.

[45] Harris RE, Zang EA, Anderson JI, Wynder EL. Race and sex differences in lung cancer risk associated with cigarette smoking. Int J Epidemil, 1993. 22(4): p.592-9.

[46] Pope M, Ashley MJ, Ferrence R. The carcinogenic and toxic effects of tobacco smoke: are women particularly susceptible? J Gend Specif Med,1999. 2(6):45-51.

[47] Medina PP, Carretero J, Ballestar E, Angulo B, Lopez-Rios F, Esteller M, Sanchez-Cespedes M. Transcriptional targets of the chromatin-remodelling factor SMARCA4/BRG1 in lung cancer cells. Hum Mol Genet. 2005;14(7): p. 973-82.

Titles and legends to figures

Figure 1. Potential lung cancer biomarkers: γ-H2AX (A,B,C), ER-β (D) and ER-α (E). A – Adenocarcinoma of the lung; B – AAH lesion (atypical adenomatous hyperplasia) and C - adjacent “normal” lung tissue. Note: Markedly increased expression of γ-H2AX with predominantly nuclear pattern of staining in female malignant/pre-malignant lung (A,B) compared to baseline expression in normal lung alveolar cells. Elevated expression of γ-H2AX in tissue from female lung tumor is correlated with expression of estrogen receptor β (ERβ) (D), focal staining in cytoplasm and nuclei of malignant cells was observed for ERα. Female, 72, IHC, PX-DAB, x40; (F) positive control – estrogen receptor β (Uterus); (G) negative control for estrogen receptor mouse IgG (lung adenocarcinoma); (H) negative control to γ-H2AX rabbit IgG (lung adenocarcinoma).

Figure 2. Processing of lung cancer images with CAPD,

A – Row image of lung tissue: Squamous cell Ca, γ-H2AX (Anti-phospho-H2A.X (ser139) rabbit Mab, Upstate Cell Signaling Solutions, Lake Placid, NY), IHC, x40;

B – showing enhanced and segmented nuclei image;

C – showing enhanced and segmented cytoplasm image;

D - Negative control to γ-H2AX (rabbit IgG), PX-DAB, x40.

Figure 3. Kaplan- Meier survival plot for the entire cohort since the surgery on lung cancer cases (n = 19). Note: survival calculated in months.

Figure 4. Identification of γ-H2AX in lung cancer and normal cell lines, and lung tumor tissue extracts. Cell extracts (30 цg/ml protein) were analyzed by SDS-PAGE, transferred to Immobilon-P transfer (PVDF) membranes and immunoblotted with antibodies to γ-H2AX at 1:5000 dilution (Anti-phospho-H2A.X (ser139) rabbit Mab, Upstate Cell Signaling Solutions, Lake Placid, NY). Relative position of the γ-H2AX on gels corresponded to known molecular size (15 kD). (A, C) γ-H2AX was enriched in samples from lung cancer cell cultures (NSCLC: A549, H23, SK-Mes 1, Calu 3), and (B) in samples from microdissected cells from lung cancer tissues. Less binding shown in samples of normal pulmonary fibroblasts (A) and dissected normal lung cells from each cancer case (B). Cell culture H69 (SCLC) show no binding (C).

Figure 5 Identification of ER-α in lung cancer/normal cells dissected from tissue. Equal amounts of cell lysate were separated on a 8% SDS-PAGE and transfered to PVDF membranes. Following blocking, membrane was treated with diluted (1:800) primary antibody for 72 hours. ER-α antibody was D-12, Santa Cruz biotechnology, CA) diluted with antibody binding buffer (1X PBS with 1% dry milk). After three washings with orbital shaking, horseradish PX-conjugated goat anti-mouse IgG (Santa Cruz) at 1:5000 dilution was added as secondary antibody for 1 hr at RT. The immunoreactive peptide at relative position 66 kD was detected by ECL Plus Western Blotting Detection Reagents (Amersham Biosciences, Piscataway, NJ). ER-α was enriched in tumor samples, compared to normal lung. **Note:** Case T629- Female (stage 1 at the diagnosis; status-Dead); Case T1133- Male (status – Alive). History of smoking for both patients is similar: mild smokers, <1 PPD.

An adaptive image segmentation process for the classification of lung biopsy images

Daniel W. McKee[a], Walker H. Land, Jr.[b], Tatyana Zhukov[c], Dansheng Song[c], Wei Qian[c]
[a]Dept. of Computer Science, Binghamton University, PO Box 6000, Binghamton, NY 13902;
[b]Dept. of Bioengineering and Biomedical Engineering,
Binghamton University, PO Box 6000, Binghamton, NY 13902;
[c]Dept. of Interdisciplinary Oncology, Colleges of Medicine
and H. Lee Moffitt Cancer Center & Research Institute,
University of South Florida, 12902 Magnolia Dr., Tampa, FL 33612

ABSTRACT

The purpose of this study was to develop a computer-based second opinion diagnostic tool that could read microscope images of lung tissue and classify the tissue sample as normal or cancerous. This problem can be broken down into three areas: segmentation, feature extraction and measurement, and classification. We introduce a kernel-based extension of fuzzy c-means to provide a coarse initial segmentation, with heuristically-based mechanisms to improve the accuracy of the segmentation. The segmented image is then processed to extract and quantify features. Finally, the measured features are used by a Support Vector Machine (SVM) to classify the tissue sample. The performance of this approach was tested using a database of 85 images collected at the Moffitt Cancer Center and Research Institute. These images represent a wide variety of normal lung tissue samples, as well as multiple types of lung cancer. When used with a subset of the data containing images from the normal and adenocarcinoma classes, we were able to correctly classify 78% of the images, with a ROC A_Z of 0.758.

Keywords: computer-assisted diagnosis, segmentation, lung cancer, support vector machines, clustering, image processing

1. INTRODUCTION

The purpose of this study was to develop a second opinion diagnostic tool that could read microscope images of lung tissue and classify the tissue sample as normal or cancerous, and if cancerous, identify the type of cancer. In order to be useful, such a tool needs to be fast (providing its classifications in minutes or seconds, rather than hours or days) as well as accurate. It should also be able to accommodate minor variations in the image due to small deviations in the sample collection and preparation process.

The problem can be broken down into three areas: segmentation, feature extraction and measurement, and classification. Segmentation is the process of separating the image into its constituent parts, isolating objects of interest. In this case, the images were divided into three parts—nuclei, cytoplasm, and background. Once the image has been accurately segmented, we can extract a great deal of information. For example, by comparing the area of the nuclei in the image with the area of the cytoplasm in the image, we can calculate the nucleocytoplasmic ratio. We can also collect a number of additional statistics throughout the image. Finally, the collected information extracted from each image can be used with a computational intelligence tool such as an SVM or Kernel-Partial Least Squares (K-PLS) to classify the image as normal or cancerous. Each of these sub-problems will be discussed in detail in Section 2.

2. METHOD

2.1 Segmentation

Each source image shows the field of view of the microscope, typically showing several hundred cells. The purpose of the segmentation process is to separate the portions of the image that correspond to nuclei, the portions that correspond to cytoplasm, and the portions which are not of interest (these may be "empty" or may include cytoplasm or other tissue not of interest). The specifics of the collection process and the image formats are discussed formally in section 3.

The segmentation process is performed in several phases. The first phase attempts to perform a coarse segmentation. Each of the remaining phases attempts to improve on this initial segmentation in some way.

2.1.1 Phase 1—Fuzzy C-Means with an adaptive kernel

We used the well-known fuzzy c-means (FCM) clustering algorithm as a starting point for our segmentation process. This algorithm uses an iterative process to divide the pixels in the source image into an arbitrary number of classes by solving a constrained optimization problem. We want to minimize

$$J(\mathbf{U},\mathbf{v}) = \sum_{k=1}^{n}\sum_{i=1}^{c}\left(u_{ki}\right)^2\left(d_{ki}\right)^2 \tag{1}$$

subject to

$$u_{ki} \in [0,1] \quad \forall k, i, \tag{2}$$

$$\sum_{i=1}^{c} u_{ki} = 1 \ \ \forall k, \tag{3}$$

and

$$0 < \sum_{k=1}^{n} u_{ki} < n \ \ \forall i. \tag{4}$$

In Eq. (1), u_{ki} is the degree of membership of pixel k in cluster i, d_{ki} is the "distance" between pixel k and the centroid for cluster i, n is the number of pixels in the image, c is the number of clusters (classes), **U** is an n x c matrix of the membership values for each pixel and cluster, and **v** is the collection of centroids, one for each cluster. The constraints listed ensure that each pixel's membership in each cluster is bound by [0,1] (Eq. (2)), and that the membership values across all clusters sum to one for each pixel (Eq. (3)). Finally, Eq. (4) ensures that each class has some pixels with a non-zero degree of membership, and that no single cluster can contain all pixels. This constraint forces the solution to have the desired number of distinct clusters.

Step 1:

Choose some scalar ? > 0 (used to detect convergence); fix integer c (the number of clusters), 2 = c < n; and initialize $\mathrm{U}^{(0)}$ (the initial membership matrix).

Step 2:

Calculate the c fuzzy cluster centers $\left\{\mathbf{v}_i^{(l)}\right\}$ using

$$v_{ji} = \frac{\sum_{k=1}^{n} {u_{ki}}^2 x_{jk}}{\sum_{k=1}^{n} {u_{ki}}^2} \quad \forall i = 1,2,\ldots,c$$

with the membership values $\mathrm{U}^{(l)}$, where l is the current iteration in the algorithm (starting with 0).

Step 3:

Update $\mathrm{U}^{(l)}$ to create $\mathrm{U}^{(l+1)}$ using

$$u_{ki} = \frac{1}{\left[\sum_{h=1}^{c}\left(\frac{d_{ki}}{d_{kh}}\right)^2\right]},$$

where d_{ki} is the distance between pixel k and the center $\mathbf{v}_i$, d_{kh} is the distance between pixel k and the center $\mathbf{v}_h$, and the values for the centers are taken from $\mathbf{v}^{(l)}$. u_{ki} is the calculated cluster membership value for pixel k in cluster i in $\mathrm{U}^{(l+1)}$.

Step 4:

Compare $\mathrm{U}^{(l)}$ and $\mathrm{U}^{(l+1)}$ using some convenient matrix norm, that is, if $\left\|\mathbf{U}^{(l+1)} - \mathbf{U}^{(l)}\right\| < x$, then stop; otherwise, l←l + 1 and return to step 2.

Figure 1: Bezdek's fuzzy c-means clustering algorithm

The optimization problem described above can be solved by Bezdek's fuzzy c-means clustering algorithm, outlned in Figure 1.[1] The initial membership matrix in step 1 can be assigned at random; however, the algorithm will converge more rapidly if we make an intelligent guess at the classification for each pixel. Once the initial membership matrix has been established, it can be used in step 2 to provide a new estimation of the position of the center of each cluster. Step 3 uses these new centers to refine the membership matrix. These two steps (using the membership matrix to refine the cluster centers, and using the centers to refine the membership matrix) are alternated until the changes to the membership matrix become insignificant.

A key element affecting the accuracy of FCM's classification is the distance measurement, d_{ki}. Bezdek defines this distance as the Euclidean norm between a sample $\mathbf{x}_k$, and a cluster center, $\mathbf{v}_i$, so that

$$d_{ki} = \left\| \mathbf{x}_k - \mathbf{v}_i \right\|. \tag{5}$$

However, the Euclidean norm may not be the most accurate representation of the "distance" between these points. By substituting a kernel function for the distance calculation in Eq. (1) yields

$$J(\mathbf{U}, \mathbf{v}) = \sum_{k=1}^{n} \sum_{i=1}^{c} \left(u_{ki} \right)^2 \left(\mathrm{K}(k, i) \right)^2 . \tag{6}$$

Now the clustering can be based on the kernel's calculation of the similarity between the sample and the cluster center. It is important to note that the distance we are referring to here is the difference in the RGB (red, green ,blue) color space between the sample pixel and the cluster center—it has nothing to do with the physical location of the sample pixel within the image. Likewise, the cluster center is the prototypical set of RGB values for a pixel belonging to that cluster—physical location within the image is not considered. In fact, FCM treats the image as a set of pixel values, without respect for the coordinate system used to assemble those pixels into a two-dimensional image.

The kernel function used in this study is based on the statistical properties of each image. The average intensity and standard deviation of the intensity are measured for each color channel (red, green, and blue), and for the image as a whole. The ratio of the color channel's standard deviation to the overall standard deviation is used to set a weight corresponding to that color channel. The rationale is that a higher standard deviation implies a broader range of values in that color channel, and therefore that channel is likely better differentiated than one with a lower standard deviation. The distance calculations then use these weights to adjust the contribution of each channel to the overall distance. For example, if the red channel had a much higher weight than the green and blue channels, a small difference in the amount of "redness" would yield a larger distance metric than the same difference in the green or blue channels.

A number of other kernels could be used, potentially dramatically improving the quality of the initial segmentation. These will be discussed in section 5.

2.1.2 Phase 2—Refining the segmentation

While the FCM segmentation is a good starting point, it does not provide sufficient accuracy to allow precise measurements of the image's features. To successfully segment the image, *a priori* knowledge about cell structure and properties needs to be taken into account and combined with information contained in the image. Yang and Jiang state that "one of the most challenging issues in medical image segmentation is to extend traditional approaches of segmentation and object classification in order to include shape information rather than merely image intensity."[2] Our refinement of the segmented image takes place in several discrete steps.

The first step involves incorporating information about neighboring pixels into the classification. Although a complete classification mechanism can be based on this neighboring pixel information,[3] our work focuses on tuning the classification made using the intensity information. For each pixel in the image, all pixels within a five-pixel radius are averaged together, using a Gaussian function to weight the average such that adjacent pixels have a high contribution to the average, while pixels four or five pixels away have a lesser contribution. By comparing the membership likelihood values for the most likely and second most likely class, we can estimate a "confidence" that the class with the highest likelihood is indeed the correct classification. This confidence value is computed for the pixel being evaluated and also for the neighborhood average. If the neighborhood average has a higher confidence then the pixel being evaluated, then the pixel being evaluated is updated by combining the original value with the neighborhood information, using the confidences as weights. This has the effect of the neighborhood "pulling" the classification of that pixel toward its average. Otherwise, if the neighborhood classification is different from that of the pixel being evaluated, the confidences are compared, and if we consider the pixel under consideration to be misclassified, its classification is simply replaced with the neighborhood information. This process is repeated through the entire image several times, until no significant changes are being made. The visual effect of this processing is similar to smoothing—as it tends to take away any very small areas which are different from their surroundings. While a certain level of smoothing is helpful in cleaning up misclassified pixels, it is possible to smooth the image to the extent where the accuracy of the original segmentation begins to suffer.

The next step attempts to find small pockets within a nucleus that have been misclassified as cytoplasm. Each pixel currently classified as cytoplasm is evaluated to see if it is surrounded by nucleus pixels. Each of eight directions is checked (N,NE,E,SE,S,SW,W,NW) to see if pixels classified as nucleus could be found within 20 pixels in that direction. If nucleus pixels are found in all eight directions, a fill routine is used to reclassify that section of cytoplasm as nucleus. If the area filled is above a certain threshold, we assume the fill process spilled into the remainder of the image and undo the fill. While this step provides solid, contiguous regions for the nuclei, it also has the potential to misclassify some cytoplasm when there are many nuclei very close together in the image.

The next step attempts to identify the approximate center of each nucleus, and also find any areas in which multiple nuclei have been joined by the segmentation process thus far. Through repeated erosion, and keeping track of when each nucleus disappears, we can find the centers. By ignoring nuclei that disappear during the first few erosions, we can eliminate very small areas that are likely misclassified. The erosion process will also separate most joined nuclei, and provide us with multiple centers in those areas. In order to be able to accurately measure cell size and shape, we need to separate any joined nuclei—this is done using a watershed line technique to draw a line of cytoplasm between centers at a point appropriate for the relative sizes of the nuclei being separated.

The final step in our segmentation process is to eliminate any excess cytoplasm not in the proximity of a nucleus. We check each cytoplasm pixel and find which nucleus it's closest to. If the distance to the nucleus is greater than the diameter of the nearest nucleus, then the pixel is reclassified as background.

While all of the post-processing is complex, it is still manageable on current PC hardware. On a 2.6GHz P4 system with 1GB of RAM, we were able to segment and extract features from a 1520x1080 image in around 10 minutes.

2.2 Feature identification and measurement

Once an acceptable automated segmentation process is in place, we can begin to take measurements from the identified regions in the image. Thiran and Macq identify some measurable nuclear features that can help identify cancerous cells. For example, the nuclei of cancerous cells are typically larger than normal cells, and vary in size (*anisonucleosis*) and shape (*nuclear deformity*). Since this increase in nucleus size is not accompanied by a corresponding increase in cytoplasm, the *nucleocytoplasmic ratio* is increased. Finally, the chromatin distribution within the nucleus is often uneven, which causes the nucleus to stain unevenly, with distinct dark and light patches (*hyperchromasia*).[4] For further details on the pathology and cytology of benign and malignant tumors, the reader is referred to Koss.[5]

Within these categories, we identified 15 metrics that could easily be extracted from the segmented image. These are presented in Table 1. Metrics 1 through 5 are straightforward descriptive statistics. During the segmentation process each nucleus center is marked. By iterating through these marked pixels, we can collect information on a nucleus by nucleus basis. Since each nucleus was separated from any adjoining nuclei during the segmentation process, we can use a simple fill algorithm to calculate the area (in pixels) of each nucleus. This information leads directly to metrics 1 and 2. The average

Metric	Description
Cell size/Anisonucleosis	
1	Average Nucleus Area
2	Standard Deviation of Nucleus Area
3	Average Cytoplasm Area
4	Average cell size
Nucleocytoplasmic ratio	
5	Nucleocytoplasmic ratio
Nuclear texture/hyperchromasia	
6	Average Nucleus Pixel Intensity (measured across entire image)
7	Standard Deviation of Nucleus Pixel Intensity (measured across entire image)
8	Average of Nucleus Average Intensity (each nucleus averaged separately)
9	Average of Standard Deviation of Nucleus Pixel Intensity (SD of each nucleus measured separately)
10	Standard Deviation Corresponding to Metric 8
11	Standard Deviation Corresponding to Metric 9
Nuclear shape/deformity	
12	Average nucleus radius (each nucleus measured separately)
13	Average of Standard Deviation of nucleus radius (SD of nucleus radius measurements measured separately for each nucleus)
14	Standard Deviation Corresponding to Metric 12
15	Standard Deviation Corresponding to Metric 13

Table 1: Descriptive metrics used to quantify image features

cytoplasm area calculation requires a different approach. Since there are no clear boundaries in the segmentation between the cytoplasm of one cell and the cytoplasm of an adjoining cell, we cannot easily measure the cytoplasm area on a cell by cell basis; however, it is fairly simple to count the number of pixels classified as cytoplasm across the entire image. Dividing this total cytoplasm area by the number of cells in the image (obtained by counting the marked centers) yields the average cytoplasm area per cell (metric 3). By adding the average nucleus area and average cytoplasm area together, we can find the average overall cell size (metric 4), and by dividing the average nucleus area by the average cytoplasm area we can find the nucleocytoplasmic ratio (metric 5).

Metrics 6 and 7 are taken from the distribution of pixel intensities for all pixels belonging to the nucleus class. We simply iterate through the nucleus pixels and calculate the average and standard deviation of the pixel intensity. For this study, the pixel intensity measure used was the average of the three color channels (that is, (R+G+B)/3). Metrics 8 through 11 require taking intensity measurements separately for each nucleus. Iterating again through the nucleus centers, and using a modified fill algorithm to identify the pixels that belong to a given nucleus, we calculate the average intensity and standard deviation of the intensity distribution *within that nucleus*. So for an image with *n* cells, we obtain a list of *n* averages and standard deviations. Metrics 7 and 9 are the average and standard deviation of the distribution of average intensities throughout all cells in the image. Metrics 8 and 10 are the average and standard deviation of the distribution of standard deviations of intensity throughout all cells in the image. So metric 6 measures how darkly the nucleus pixels stained, and metric 7 measures the homogeneity of the nucleus intensities across all nucleus pixels. Metric 8 provides a similar measure to 6, but using per pixel averaging. Metric 9 gives us a measure of the homogeneity of the intensity within each pixel (Is there light and dark within each nucleus, or are the nuclei individually homogenous in intensity, with the intensity variation being between one nucleus and the next?). Finally metric 11 tells us whether all nuclei have a similar intensity distribution, or whether some nuclei are much more homogenous than others.

The last four metrics are calculated in a similar manner to 8 through 11. We iterate through each nucleus, and measure the radius of the nucleus in eight cardinal and intermediate directions. The average and standard deviation of these radius measurements is calculated (for each nucleus), and metrics 12 through 15 calculated based on that distribution.

These metrics were collected for all the images in the data set. The raw metrics were then normalized by scaling them linearly to the range [-1,+1]. This normalized collection of metrics was used as the basis for the final step of the process, classification.

2.3 Classification using a Support Vector Machine (SVM)

A support vector machine (SVM) is a kernel-based classifier which trains to a global minimum. The classifier is "trained" by solving an optimization problem using the training data. The objective function for the SVM if given by the Lagrangian:

$$\mathbf{J}(\mathbf{w}, b, a) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{\mathbf{i=1}}^{N} a_i \left[y_i \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b \right) - 1 \right], \tag{7}$$

where the non-negative variables α_i are the Lagrange multipliers. The solution to the constrained optimization problem is determined by the saddle point of the Lagrangian function. The minimum of **J** with respect to the weight vector and bias leads to the solution for the Lagrange multipliers $\{\alpha_i\}$ for i = 1, 2, … N.

To obtain a solution in the most general case where the environment is non-linear and non-separable, two concepts are introduced into the formulation: 1) slack variables, and 2) inner-product kernel functions. Slack variables are a measure of the deviation of a sample point from the ideal condition of pattern separability. The inner-product kernel function is used to construct a decision surface that is non-linear in the input space, but its image in the high-dimensional feature space is linear. The inner product kernel function must be symmetric and must satisfy Mercer's Theorem. The solution to the Lagrangian dual problem for this most general case is given by:

$$\max \mathbf{Q}(a) = \sum_{i=1}^{?} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j y_i y_j \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) \tag{8}$$

subject to the constraints:

$$\sum_{i=1}^{N} a_i y_i = 0 \text{, and} \tag{9}$$

$$0 \le a_i \le \mathrm{C}. \tag{10}$$

Our implementation used sequential minimum optimization (SMO)[6] to solve Eq. (8). Additional references contain extensive information about the mathematical foundations and development of SVMs.[7-8]

Conceptually, the purpose of the optimization problem is to find the optimal separating hyperplane between the classes. The non-zero Lagrange multipliers define the sample points which form the class boundary. The kernel function describes the similarity (or dissimilarity) of two samples in the kernel's feature space. Choosing a kernel function and kernel parameters that accurately represent the true relationship between the points is essential to obtaining good classification performance from the SVM. A detailed discussion of kernel functions can be found in Land, et. al.[9]

2.3.1 Multi-class classification problem

In order to use the an SVM to classify the images in this data set, we must reconcile the binary classification behavior of the SVM with the fact that the images represent five distinct classes, each with its own properties. Two general approaches are possible: simplify the data set into two classes, or devise a mechanism for classifying an unknown based on a pairwise comparison with each of the five classes.

In the first approach, the individual cancer types could be replaced with a single class that represents cancers (regardless of type). This substitution reduces the data set from five classes to two classes—namely, normal and cancerous. While this simplifies the classification, mixing several types of cancer in the same class may reduce the accuracy with which we are able to identify those cancers.

If we want to limit our classification process to identifying only one type of cancer, we can train the SVM using only the normal cases (as the first class) and the particular type of cancer we're interested in identifying (as the second class). This has the advantage of potentially increasing the accuracy of detecting that particular type of cancer; however, the "appropriate" classification for an image of another type of cancer is undefined. Ultimately, we want a process that will classify an unknown case into the particular class to which it is most similar. This study evaluates the performance of the process for the normal vs. cancer and normal vs. adenocarcinoma cases. Generalizing the process to provide a robust multi-class solution is planned future work.

3. DATA

The source data used by this process is a collection of 85 lung tissue images collected from patients seen at the Moffitt Cancer Center and Research Institue. The tissue samples were collected by surgical or needle biopsy, and stained by r-H2AX, PX-DAB, and hematoxylin counterstain. A 1520x1080 pixel 24-bit color digital image of the specimen was then made through an optical microscope with a 10X eyepiece and a 40X objective lens, for a total magnification of 400X. Of the 85 images, 37 are images of normal lung tissue, 21 are images of adenocarcinomas, 15 are images of sqaumous cell carcinomas, ten are of bronchoalveolar carcinomas, and two are of atypical adenomatous hyperplasia (a pre-malignant lesion).

The field of view of the microscope may contain between approximately 200 up to 1,000 cells. Also, some specimens more rapidly acquire stain than others, and some images have cytoplasm which picked up the staining solution in widely varying intensities, yielding

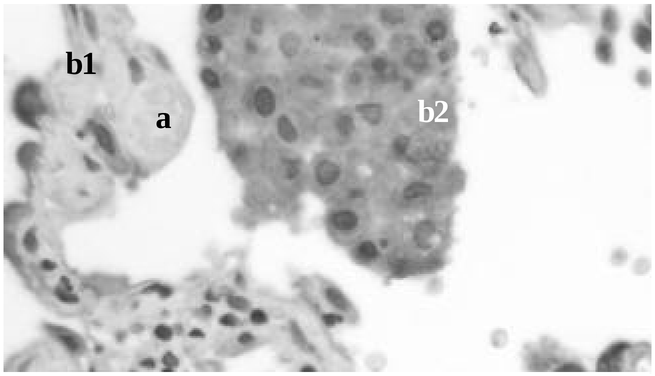


Figure 2: Cytoplasm not associated with any cell (a) and differences in cytoplasm stain intensity (b1, b2)

some areas of “dark” cytoplasm much closer in color to the nuclei than to the remaining cytoplasm. Finally, many images also have “noise” present in the form of cytoplasm which isn’t associated with a cell in the image. This artifact of the microscope slide preparation process needs to be treated as background, even though it has the same appearance and intensity as the cytoplasm surrounding each cell. Figure 2 shows a portion of one of the images illustrating some of these challenges. Since the image properties vary substantially from one image to the next, the segmentation process must adapt to the image at hand in order to accurately segment each image.

Figure 3(a) shows the field of view for a non-cancerous sample, and figure 3(b) shows the full resolution detail of the marked are from the sample image. Figure 4 shows a similar pair for an adenocarcinoma image. From the detailed area, it is clear that this image provides more of a challenge for the segmentation process, as the nuclei and cytoplasm have nearly identical intensities.

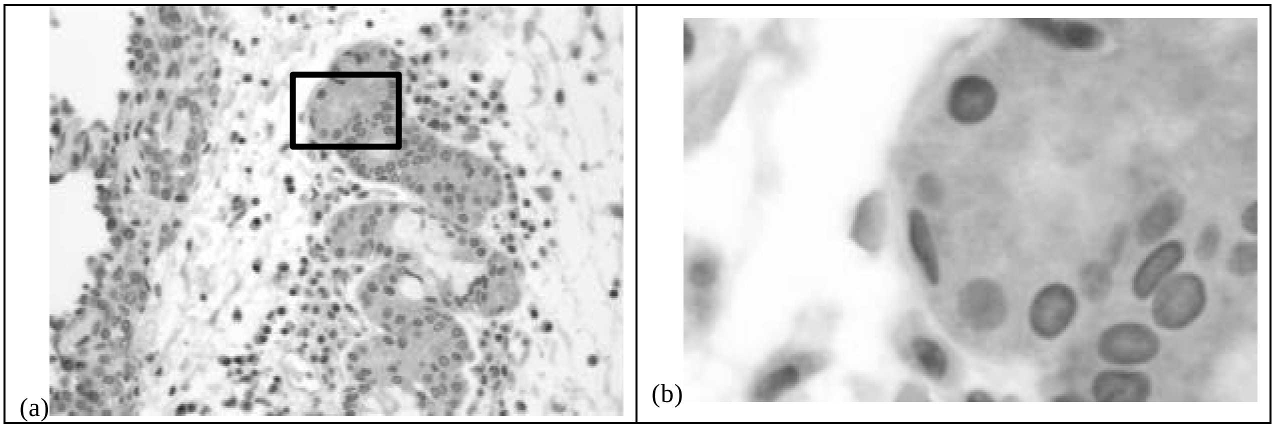


Figure 3: Microscope field of view for a normal tissue sample image (a), and full resolution view (b) of marked area

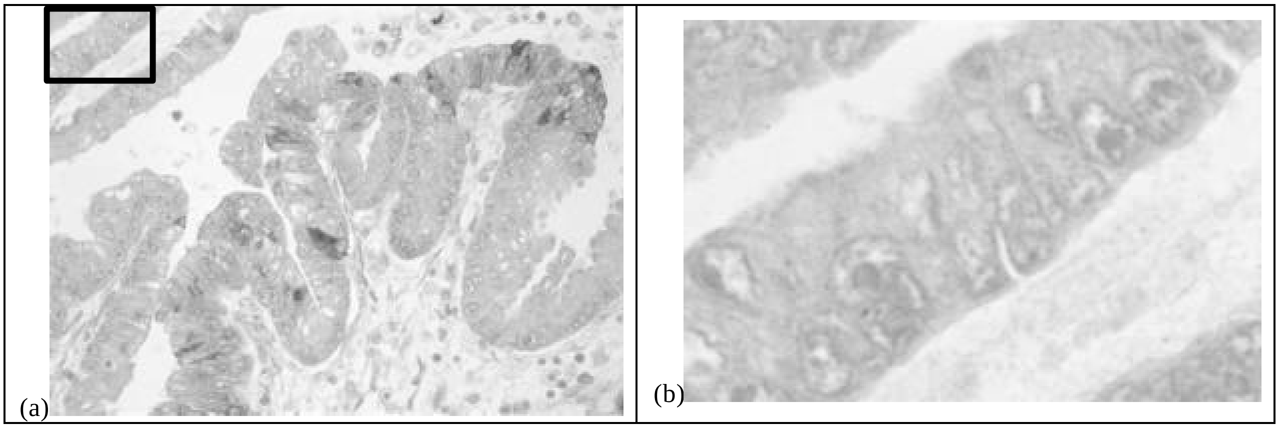


Figure 4: Microscope field of view for an adenocarcinoma sample image (a), and full resolution view (b) of marked area

4. RESULTS

The results of this study will be presented in three parts. First, we will show examples of images segmented by this process. Next we will provide examples of feature measurements for some of the images. Finally, we present the classification results when the extracted metrics are used to predict the classification for each image.

4.1 Segmentation results

Figures 5 and 6 show the segmentation results on the images from figure 3 and 4. In each case, (a) is a reduction of the segmentation of the entire image, (b) is a mask showing only the nuclei from the original image (using the same enlarged area as the earlier figures), and (c) is a mask showing only the cytoplasm from the original image in the same area. A close examination of these segmentations illustrates several steps in the segmentation process. The three solid arrows in Figure 5(b) show watershed lines where joined nuclei were separated. The two on the right have near-perfect placement of the separation. The one on the left cuts through the middle of a nucleus—this was caused by the erosion process only identifying two nucleus centers in what was actually three joined nuclei (the nucleus with the line through it, the one above and the one below are all connected). While not ideal, this misplaced separator should have minimal impact when data from all of the cells is averaged together. Also in Figure 5(b), the dashed arrow shows a pair of nuclei mistakenly classified as one—once again, the erosion process resulted in only one center for this area, but since we area averaging data together for hundreds of cells, the impact of occasional anomalies is minimized. The excess cytoplasm cleanup can be seen in Figure 5(c). The reader is referred back to Figure 3(b) to see that the right half of the image has large areas of cytoplasm not near any of the nuclei. In Figure 5(c) the limiting of the cytoplasm under consideration to that which in near an identified nucleus can be clearly seen.

Figure 6(a) shows a complete segmentation for the adenocarcinoma image from figure 4. While much of this segmentation is surprisingly good considering the limited contrast between the nuclei and cytoplasm in many areas (see Figure 4(b)), the area in the upper right of Figure 4(a) shows a large solid area classified as nuclei. The erosion process attempts to break this large area into individual "cells," but the result has no real correlation to the underlying image data in that part of the image. Identifying areas such as this and removing them (by resetting their classification to "background") could possibly improve the accuracy of the statistics collected and is an area of ongoing investigation. The remainder of the image seems to be segmented well, as can be seen in the nuclei and cytoplasm masks in Figure 6(b) and (c).

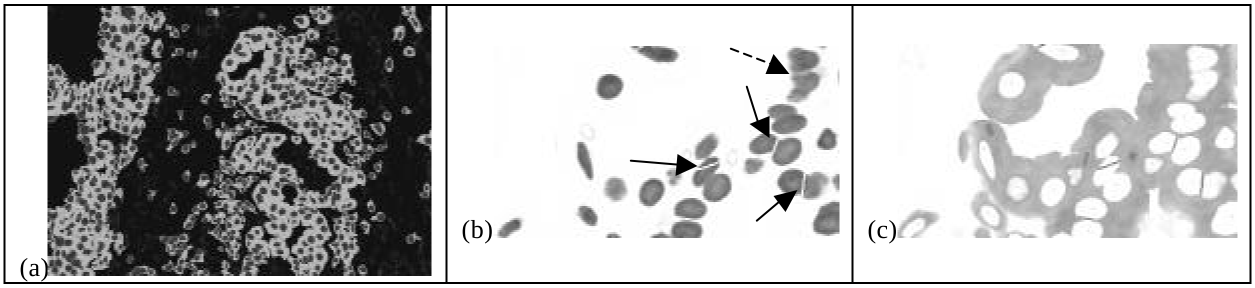


Figure 5: Segmentation of the image from figure 3

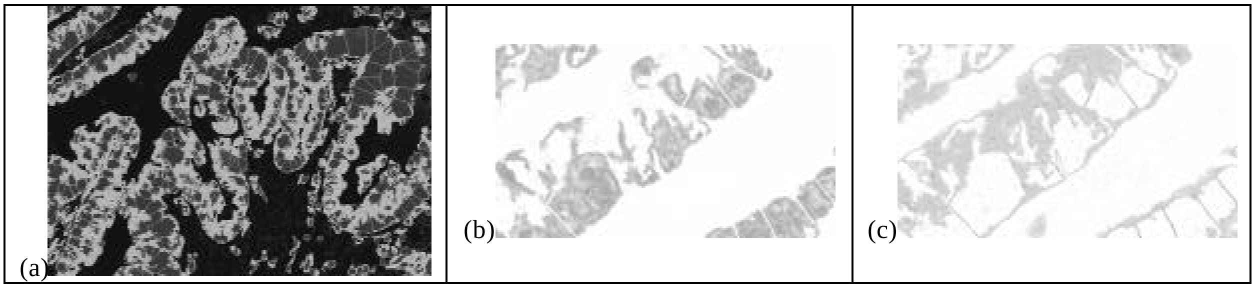


Figure 6: Segmentation of the image from figure 4

4.2 Feature measurement

The metrics listed in Table 1 were collected for each of the 85 segmented images. The metrics corresponding to the images in Figure 5 and 6 are summarized in Table 2. It is clear that the accuracy of these measurements depends largely on the accuracy of the segmentation process; however the averaging of data across a large number of nuclei can compensate for minor discrepancies in the segmentation or measurement of a small portion of the image.

4.3 Classification

The final goal of this work was to accurately classify the images as normal or cancerous. The features measured were all scaled as described in section 2.2, and tested using a support vector machine and one-hold-out cross-validation. This cross-validation technique allows us to make the most of this limited collection of data. Each case is "held out" one at a

Metric	Description	Figure 3	Figure 4
Cell size/Anisonucleosis			
1	Average Nucleus Area	478.5	1089.7
2	Standard Deviation of Nucleus Area	259.2	1156.2
3	Average Cytoplasm Area	1067.2	1525.5
4	Average cell size	1545.7	2615.2
Nucleocytoplasmic ratio			
5	Nucleocytoplasmic ratio	0.448 (1:2.2)	0.714 (1:1.4)
Nuclear texture/hyperchromasia			
6	Average Nucleus Pixel Intensity (measured across entire image)	136.2	184.0
7	Standard Deviation of Nucleus Pixel Intensity (measured across entire image)	19.1	16.4
8	Average of Nucleus Average Intensity (each nucleus averaged separately)	137.4	186.0
9	Average of Standard Deviation of Nucleus Pixel Intensity (SD of each nucleus measured separately)	14.2	9.9
10	Standard Deviation Corresponding to Metric 8	11.0	8.8
11	Standard Deviation Corresponding to Metric 9	5.8	4.3
Nuclear shape/deformity			
12	Average nucleus radius (each nucleus measured separately)	11.0	14.8
13	Average of Standard Deviation of nucleus radius (SD of nucleus radius measurements measured separately for each nucleus)	2.6	4.9
14	Standard Deviation Corresponding to Metric 12	2.9	7.5
15	Standard Deviation Corresponding to Metric 13	1.7	3.5

Table 2: Sample image feature measurements

time from the data set. The remainder of the set is used to train the SVM (by solving the optimization problem described in section 2.3), and the SVM's prediction for the held out case is recorded. The next case is then held out and a new SVM trained, and the process is repeated until all cases have been used.

Eighteen trial runs were completed using four different kernels, and varying the parameters for each of the kernels. This process was not an exhaustive search to find the kernel best suited to this problem, but rather an attempt to find a "good" solution, and determine the feasibility of using the collected data to classify the samples. While the results listed leave room for improvement, they also demonstrate the potential for an automated end-to-end process such as this to work well.

When using the entire set of 85 images, the normal tissue samples were placed in one class, and all of the other (cancerous) images were placed in the other class. The SVM provided an output between -1 and +1. For the purposes of determining how many samples were correctly classified, we placed an arbitrary threshold at zero. We also performed an ROC analysis in order to estimate classification performance over all possible threshold settings. The best performing SVM configurations used a kernel based on the Euclidean distances between samples. Further details on this non-linear kernel can be found in Schölkopf.[10] We were able to correctly classify 51 of the 85 samples—a classification accuracy of 60%. Our ROC Analysis yielded an A_Z of 0.623. While these results are not exceptional, they are significantly better than random. Further discussion of possible improvements is included in section 5.

A second battery of tests was done using only the normal and adenocarcinoma image data. In this case, all of the normal images were placed in one class, and all of the adenocarcinoma images in the other class. Because the cancer class contains only one type of cancer, we expect the class to be more clearly defined than in the case where several types of cancer were mixed in the same class. A similar set of eighteen trial runs was completed using one-hold-out cross-validation with this subset of 58 images. The best performing kernel in this set was a radial basis function kernel. It correctly classified (using the same standard as above) 45 out of the 58 samples, yielding an accuracy of 78%. The ROC Analysis yielded an A_Z of 0.758, much better than random, though still with room for improvement.

5. CONCLUSIONS

The prototype developed works well and is fairly robust in its ability to adapt to different images without user intervention. A number of further refinements are possible in each step of the process.

Image segmentation accuracy could be improved by use of a more intelligent distance measure during the initial FCM segmentation (such as one that accounts for gamma and maps the color space into a more linear representation of perceived brightness). Adding heuristics to identify and eliminate large irrelevant or "unclassifiable" areas (such as the upper right corner of the image in figure 6) could also lead to improved accuracy when measurements are taken, or at least provide a warning when the results are questionable. Additional feature measurements could extract more information from the segmented image—the metrics collected were chosen because they were easier to quantify and collect from the image data; however, a large number of other features exist and are used by cytologists to help evaluate samples. Certainly some of these features could be measured analytically from the image data.

Further tuning of the SVM and the multiple-class evaluation process is needed to maximize performance. Other classification mechanisms should be evaluated as well, such as partial least squares (PLS) or Kernel-Partial Least Squares (K-PLS). A PNN such as in Morrison and Attikiouzel[3] could provide a possible alternate or supplemental initial segmentation technique for images where the current process has difficulty.

In spite of the rather numerous possible improvements, the results support the idea that an automated end-to-end image classification process is possible. In particular, the results using a single type of cancer are very good considering the noted shortcomings of the current segmentation process and how little optimization was done to the classifier.

ACKNOWLEDGEMENTS

This work was supported by the efforts of a number of graduate and undergraduate students. The authors particularly wish to acknowledge the contributions of Cuong Dang and Jason Cole to the success of this project.

REFERENCES

1. Bezdek, J. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.
2. Yang, F. and Jiang, T. "Cell Image Segmentation with Kernel-Based Dynamic Clustering and an Ellipsoidal Cell Shape Model." *Journal of Biomedical Informatics* **34** (2001), pp. 67-73.
3. Morrison, M. and Attikiouzel, Y. "A Probabilistic Neural Network Based Image Segmentation Network for Magnetic Resonance Images." *Proceedings of the International Joint Conference on Neural Networks* **3** (1992), pp. 60-65.
4. Thiran, J. and Macq, B. "Morphological Feature Extraction for the Classification of Digital Images of Cancerous Tissues." *IEEE Transactions on Biomedical Engineering* **43**:10 (October 1996), pp. 1011-1020.
5. Koss, L. *Diagnostic Cytology and Its Histopathologic Bases*. Lippincott, Philadelphia, PA, 1992.
6. Platt, J. "Fast training of Support Vector Machines Using Sequential Minimal Optimization", in B. Scholkopf, C.J.C. Burges, and A.J. Smola, Eds., *Advances in Kernel Methods: Support Vector Learning*, MIT Press, 1998.
7. Cristanini, N., and Shawe-Taylor, J., *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods*, Cambridge University, 2000.
8. Burges, C.J.C. "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, **2**:2, pp. 121-167, Kluwer, Boston, 1998.
9. Land, W. H., Jr., McKee, D. W., Lo, J. Y., & Anderson, F. R. "Improving mammogram screening using a bank of support vector machines (SVMs)." *Artificial Neural Networks in Engineering (ANNIE '02)*. St. Louis, MO, 10-13 November, 2002.
10. Schölkopf, Bernhard. *The Kernel Trick for Distances*. Microsoft Technical Report MSR-TR-2000-51. 19 May 2000.